Bridging the Modality Bottleneck in Pathology MIL through Molecularly Guided Feature Adaptation

Yucheng Xing[1,2*] **Pei Liu**[3*] **Jingying Ma**[1] **Ruping Hong**[4] **Jiangdong Qiu**[4]

Tianyu Liu[1] **Kai He**[1] **Ling Huang**[1,5†] **Mengling Feng**[1]

[1]National University of Singapore
[2]National University of Singapore Guangzhou Research Translation and Innovation Institute
[3]Hunan University
[4]Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College
[5]Imperial College London

Abstract

Multiple instance learning (MIL) is the dominant framework for whole-slide image analysis in computational pathology, typically combining a frozen patch encoder, a projection layer, and a slide-level aggregator. While encoders and aggregators have been extensively studied, the projection layer remains a largely morphology-only bottleneck. This limits endpoints such as biomarker status and survival, which are governed by a molecular state that is not fully captured by H&E morphology. We introduce Molecularly Informed Spatial-Transcriptomic Transform (MIST), a plug-in replacement for the MIL projection layer that uses paired spatial transcriptomics only during training to construct molecular views. MIST clusters gene expression profiles into cross-modal prototypes, anchors them in the frozen foundation model feature space, and uses them to reorganize H&E patch features according to molecular programs. It requires no transcriptomics at inference and can be inserted before standard MIL aggregators. We evaluate MIST across 23 downstream tasks and 8 MIL aggregators. MIST improves 240 of 256 configurations over the standard projection layer, with an average gain of +3.5 points, observed consistently across endpoint types: +5.2 points on survival prediction, +3.3 points on tissue subtyping, and +2.1 points on biomarker prediction. Ablations show that gene-derived prototypes contribute beyond morphology-only alternatives, while spatial, biological, and pathological analyses show that cross-modal prototype affinities capture spatially coherent molecular programs from H&E alone. Code is available at `https://github.com/YuchengXing99/MIST`.

1 Introduction

Computational pathology (CPath) increasingly relies on whole-slide image (WSI) analysis for tissue subtyping, biomarker prediction, and prognosis [1, 2, 3]. Because WSIs are gigapixel-scale and heterogeneous, multiple instance learning (MIL) has become the dominant paradigm [4]. A slide is split into patches, each embedded by a frozen pathology foundation model (FM), and the resulting patch set is aggregated into a slide-level prediction. This pipeline has been effective, but it exposes a fundamental mismatch. Many clinical endpoints, including biomarker status, treatment response,

*Equal contribution.
†Corresponding author: Ling Huang (`iweisskohl@gmail.com`).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and survival, are governed by molecular state, whereas frozen pathology FMs are pretrained mainly from H&E morphology and pathology reports [5]. They organize tissue along morphological axes. Histologically similar regions can correspond to distinct molecular programs (Fig. 1A), suggesting that morphology-only MIL faces not only a capacity bottleneck [6], but also a *modality bottleneck*: the projection layer lacks molecular supervision to organize H&E features [7, 8, 9].

We address this bottleneck at the projection layer. A standard MIL pipeline contains a frozen patch encoder, a per-patch projection layer, and a slide-level aggregator [6]. Prior work has extensively improved encoders through self-supervised and vision-language pretraining [10, 11], and aggregators through attention, clustering, and low-rank structure [12, 13, 14]. The projection layer is a strategically important intervention point because it can reshape patch features while preserving both the frozen encoder and the downstream aggregator. Existing projection-layer redesigns still operate inside the FM's morphological feature space [15]. They can better use morphology, but receive no molecular supervision on how to organize it.

Pathologists routinely overcome this limitation by adding molecularly targeted stains. When H&E morphology alone cannot resolve a diagnostic or prognostic question, immunohistochemistry (IHC) allows the same tissue to be interpreted under an additional molecular label space [16]. Inspired by this role of staining, we introduce Molecularly Informed Spatial-Transcriptomic Transform (MIST) for MIL. MIST converts paired spatial transcriptomics (ST), available only during training, into molecular views that reorganize H&E patch features along gene-defined axes. It leaves the pathology FM frozen, requires no transcriptomics at inference, and can be inserted before standard MIL aggregators.

Paired ST is a natural source of molecular supervision because it co-registers H&E morphology with spot-level gene expression [17]. Existing training-only ST approaches usually inject this signal at the encoder stage by fine-tuning or re-pretraining a pathology FM with vision-omics objectives [7, 18]. This validates the value of molecular supervision, but ties the resulting representation to a specific FM checkpoint and requires costly retraining when new pathology FMs are adopted. MIST instead moves molecular supervision one stage later. By operating at the projection layer, it provides an encoder-agnostic, molecularly informed adaptation module without modifying the frozen encoder. MIST is a form of cross-modal adaptation for frozen foundation-model pipelines. Instead of requiring the auxiliary modality at inference, MIST uses it during training to define a molecularly grounded coordinate system. The downstream MIL objective then learns how to use this coordinate system for each endpoint.

MIST realizes feature-level molecular views in three steps. First, it derives molecular prototypes by clustering gene-expression profiles and anchoring the resulting prototypes in the pathology FM feature space, allowing each prototype to encode the morphological signature of a distinct molecular concept. Second, each patch is softly associated with multiple prototypes through affinities, reflecting the fact that local tissue regions often express overlapping molecular programs. Finally, prototype-conditioned low-rank residual transforms generate one molecular view per molecular concept, which are combined according to the affinity weights and added back to the original patch embedding, preserving the underlying H&E morphology as a stable reference for downstream MIL.

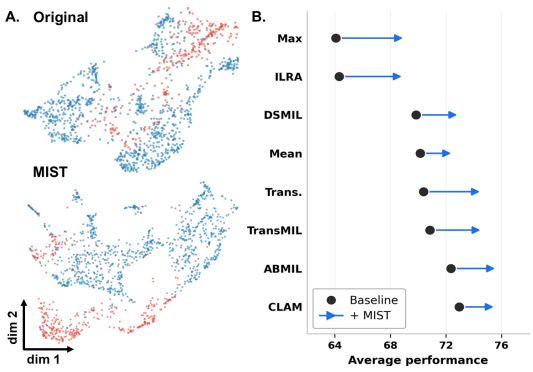


Figure 1: **(A)** MIST separates patches with similar morphology but distinct molecular profiles. Each point is an H&E patch from HEST-1k, shown before and after MIST and colored by its matched molecular cluster. **(B)** Mean performance gains across downstream tasks for eight MIL aggregators.

We evaluate MIST across 23 downstream tasks and 8 MIL aggregators using a frozen encoder. MIST improves 240 of 256 configurations over the standard projection layer, with consistent gains across endpoint types, including survival prediction (+5.2 points), tissue subtyping (+3.3 points), and biomarker prediction (+2.1 points), yielding an overall average improvement of +3.5

points. Ablation studies further show that replacing gene-derived prototypes with random or morphology-derived alternatives consistently reduces performance, suggesting that cross-modal molecular priors contribute beyond additional morphology-only modeling capacity. Our contributions are as follows:

- **A modality-bottleneck view of MIL projection layers.** We identify the projection layer as a modular but modality-limited stage in frozen-FM pathology MIL, and show that molecular priors provide an effective intervention complementary to morphology-only projection redesigns.
- **Molecular view through cross-modal prototypes.** We introduce a projection-layer mechanism that injects spatial transcriptomic priors during training while requiring only H&E at inference, enabling molecularly informed adaptation without retraining the pathology FM.
- **Broad empirical evidence and molecular interpretability.** Across 23 tasks and 8 MIL aggregators, MIST improves most evaluated configurations (Fig. 1B), while spatial, biological, and pathological analyses indicate that cross-modal prototype affinities capture spatially coherent and pathologically meaningful molecular programs.

2 Related Work

MIL for Computational Pathology. The standard MIL pipeline for gigapixel WSI analysis decomposes into three stages: patch feature extraction, per-patch linear projection, and aggregation into a slide representation. Existing work focuses on the first and third stages. At the encoder level, pathology foundation models (FMs) have evolved along two directions distinguished by pretraining supervision: vision-only FMs learn patch representations through self-supervision on unlabeled H&E [10, 19, 20, 21], while vision-language FMs additionally align H&E with pathology reports through contrastive objectives [11, 22, 23]. At the aggregation level, MIL methods pool patch sets through inductive biases such as attention [24], clustering-constrained attention with instance supervision [13], dual-stream max-pooling [25], correlated self-attention [12], and low-rank structure [14].

The intermediate projection layer has only recently received dedicated attention. Regional Transformers re-embed patches based on spatial proximity [15], query-aware sparse attention extends this to longer-range spatial dependencies [26], and low-rank mixtures of experts adapt the per-patch transformation to each patch's morphological phenotype [6]. These methods reorganize patches along progressively finer dimensions, moving from spatial to morphological similarity, yet all draw their inductive bias from the FM's feature space alone. MIST instead injects transcriptional prior from paired ST, introducing supervision from a modality the FM never saw, with no ST required at inference.

Spatial Transcriptomics for Downstream Pathology Tasks. Recent studies have explored paired molecular and H&E data for downstream MIL tasks such as tissue subtyping, biomarker prediction, and survival analysis. Multimodal fusion approaches combine H&E features with bulk molecular profiles [3, 27], but require gene expression at inference, which is rarely available in clinical practice. ST offers a finer-grained alternative, providing molecular signal co-localized with morphology at spot resolution. Training-only ST approaches incorporate this signal at the encoder stage through paired-ST objectives such as vision-omics contrastive fine-tuning [7], ST-token distillation under DINO self-supervision [8], and vision-ST mixture-of-experts pretraining on HEST-1k [9]. These studies demonstrate the value of training-time molecular supervision, but tightly couple the learned representations to a specific FM checkpoint. Given the rapid evolution of pathology FMs [10, 11], the costly vision-omics training must be repeated whenever a new FM is adopted. In contrast, MIST operates at the projection layer above any frozen FM, so adopting a new FM only requires rebuilding the prototype bank.

Prototype-Based Methods in Computational Pathology. Prototype-based methods learn a compact set of representative concepts to summarize complex inputs [28, 29, 30, 31, 32]. In CPath, prior work models patches as Gaussian mixtures and aggregates through the mixture parameters [33], clusters patches into histological types and represents slides as type-frequency distributions [34], extends prototype-based representations to multimodal survival prediction with bulk transcriptomics at both training and inference [35], and couples morphological prototypes with class anchors under an evidential fusion framework for uncertainty-aware survival [36]. In all of these methods, prototypes live in the FM's morphology feature space and act at the aggregation stage. MIST departs in two

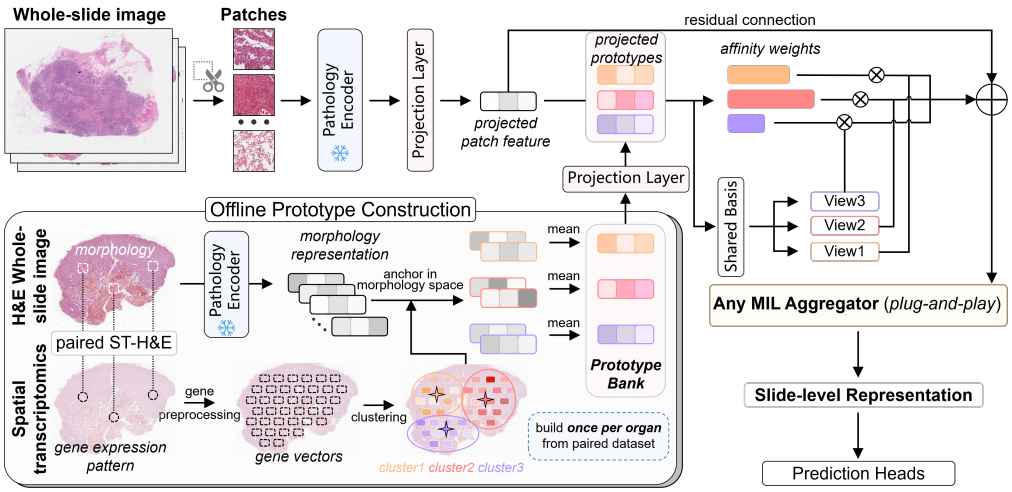


Figure 2: **Overview of MIST.** Each patch is encoded by a frozen foundation model and matched to organ-specific molecular prototypes through soft affinities. Prototype-conditioned low-rank residual transforms generate molecular views over a shared basis, which are combined to produce a patch representation consumable by any MIL aggregator. The prototype bank is constructed once per organ from paired ST and H&E data, with no transcriptomics required at inference. Each prototype $\boldsymbol{\mu}_k \in \mathbb{R}^{1536}$ (bank $K \times 1536$) is projected to $\mathbf{p}_k = \phi(\boldsymbol{\mu}_k)$.

coupled ways: prototypes are discovered in gene-expression space but anchored as centroids in morphology space, encoding the morphological signatures of molecular concepts; and they condition the per-patch projection rather than summarize the slide, leaving the aggregator unconstrained.

3 Method

To bridge the transcriptional modality gap without requiring ST at inference, we propose **MIST** (**M**olecularly **I**nformed **S**patial-**T**ranscriptomic **T**ransform), a plug-and-play replacement for the standard MIL projection layer that, unlike prior MIL redesigns, conditions patch transforms on gene-expression prototypes constructed offline from paired ST data (Fig. 2). MIST (1) builds molecular prototypes via gene-space clustering, anchored in morphological feature space for direct H&E conditioning (§3.1); (2) assigns patches via soft, non-competitive affinities to handle co-expression (§3.2); (3) applies a prototype-conditioned low-rank transform over a shared morphological subspace to produce molecular patch views (§3.3); and (4) mixes views residually by affinity, preserving the original morphological signal as a stable anchor (§3.4).

3.1 Offline Cross-Modal Prototype Construction from Paired ST-H&E Data

We derive the prototypes from paired ST and H&E data in HEST-1k [17], which provides gene expression and morphology co-localized at Visium-spot resolution across 1,276 paired slides and 26 organ types with millions of spots.

Extracting Paired Morphology and Gene Expression at Each Spot. HEST-1k provides pre-extracted patches at $0.5\,\mu$m/px, each centered on a Visium spot. We encode each patch with the frozen pathology foundation model UNI2-h [10] to obtain its morphology embedding $\mathbf{x}_j \in \mathbb{R}^D$. For the corresponding Visium spots, we follow SEAL's processing pipeline [7] to obtain the gene expression vector $\mathbf{g}_j$: empty spots and genes detected in less than 10% of spots are removed; raw counts are normalized to 10^4 per spot and $\log(1+x)$-transformed; the top 2,000 highly variable genes are retained via the Seurat v3 method [37]; and each spot's expression is averaged with the mean of its immediate neighbors on the hexagonal lattice to mitigate Visium sparsity and noise. This yields paired samples $\{(\mathbf{x}_j, \mathbf{g}_j)\}_{j=1}^{M}$. Further details are provided in Appendix A.

Clustering Within Each Organ to Form Prototypes. We construct prototypes per organ since molecular concepts differ substantially across organs and downstream tasks are typically organ-

specific. For each organ o, we cluster its gene vectors $\{\mathbf{g}_j\}$ into K_o groups (selected by silhouette score), assigning each spot a cluster label c_j. The prototype for cluster k is the centroid of the corresponding patch features:

$$\boldsymbol{\mu}_k = \frac{1}{|\mathcal{C}_k|} \sum_{j \in \mathcal{C}_k} \mathbf{x}_j \in \mathbb{R}^D, \qquad \mathcal{C}_k = \{j : c_j = k\}. \tag{1}$$

This construction is central to MIST: clustering in gene space discovers molecular concepts, while the prototypes themselves live in morphological feature space so they can condition H&E patch embeddings directly. Each $\boldsymbol{\mu}_k$ thus encodes the morphological signature of a specific molecular concept, and defines the corresponding molecular view applied to patches downstream.

3.2 Patch Assignment to Multiple Molecular Concepts

Given the prototypes $\{\boldsymbol{\mu}_k\}_{k=1}^K$, we associate each patch in a WSI with these molecular concepts. To compute the affinity of a patch to each prototype, we first project both into a shared d-dimensional space (d=512) via a linear layer ϕ, yielding $\mathbf{h}_i = \phi(\mathbf{x}_i)$ and $\mathbf{p}_k = \phi(\boldsymbol{\mu}_k)$. The affinity of patch i to prototype k is then

$$\alpha_{ik} = \frac{\sigma(\mathbf{h}_i^\top \mathbf{p}_k)}{\sum_{k'} \sigma(\mathbf{h}_i^\top \mathbf{p}_{k'})}, \tag{2}$$

where σ is the sigmoid, allowing a patch to activate multiple prototypes simultaneously when it expresses several molecular concepts at once. This reflects a common situation in tissue, where a single spot can carry mixed cellular programs. The ℓ_1 normalization preserves the relative pattern of co-activations while bounding the total affinity of each patch to a fixed budget, keeping the view intensity comparable across patches before aggregation.

3.3 Prototype-Conditioned Low-Rank Transforms

To represent how each patch manifests under each molecular concept, each prototype k induces its own view $\mathbf{v}_{ik}$ of the patch embedding, computed as

$$\mathbf{v}_{ik} = \mathbf{h}_i + \mathbf{U}_k \mathbf{V} \mathbf{h}_i, \tag{3}$$

with $\mathbf{V} \in \mathbb{R}^{r \times d}$ shared across prototypes and $\mathbf{U}_k \in \mathbb{R}^{d \times r}$ prototype-specific (r=16). The residual form mirrors the physical nature of staining. A stained sample still contains the underlying tissue, and the stain is added on top. The factorization further restricts this stain. Morphology has only a limited number of axes along which any stain can act, so the shared $\mathbf{V}$ extracts these common axes while each $\mathbf{U}_k$ recombines them for its own concept.

3.4 Patch Representation for Downstream MIL

Given the K views $\{\mathbf{v}_{ik}\}_{k=1}^K$, we combine them into a single patch embedding that the downstream aggregator can consume. Each patch is transformed by the concepts it expresses, weighted by its affinity,

$$\tilde{\mathbf{h}}_i = \mathbf{h}_i + \sum_{k=1}^{K} \alpha_{ik}\, \mathbf{v}_{ik}. \tag{4}$$

The affinity weights give each patch its own view mixture on a comparable scale across patches. Adding $\mathbf{h}_i$ once more keeps the original view of the patch available alongside the molecular views, so the original morphological signal from the frozen encoder remains accessible regardless of how the patch aligns with the prototypes.

Theoretical grounding. MIST can amplify the separation between morphologically similar patches that receive different prototype affinities. For two patches with $\mathbf{h}_i - \mathbf{h}_j = \epsilon \mathbf{v}$ but different prototype affinities, the output difference decomposes as

$$\tilde{\mathbf{h}}_i - \tilde{\mathbf{h}}_j \;=\; (2\mathbf{I} + \bar{\mathbf{W}})\, \epsilon \mathbf{v} \;+\; \Big(\sum_k \Delta\alpha_k\, \mathbf{U}_k\Big) \mathbf{V} \bar{\mathbf{h}}, \tag{5}$$

where $\epsilon \mathbf{v}$ denotes the small morphological difference between the two patches, with ϵ controlling the perturbation magnitude and $\mathbf{v}$ the perturbation direction. The first term corresponds to the

morphology-only transformation, while the second term captures molecular routing induced by differences in prototype affinity. Here, $\bar{\mathbf{W}}$ denotes the uniform-affinity baseline transform and $\Delta\alpha_k = \alpha_k(\mathbf{h}_i) - \alpha_k(\mathbf{h}_j)$ measures the affinity difference between the two patches. When the routing term aligns with the morphology term, the output distance exceeds the morphology-only baseline. A formal existence proof is provided in Appendix B.

4 Experiments

4.1 Datasets

Tissue subtyping. We evaluated MIST on six tissue subtyping classification tasks: EBRAINS fine-grained (EBRAINS-F, C=30) and coarse-grained (EBRAINS-C, C=12) brain tumor subtyping (n=2,319 slides) [38]; Non-Small Cell Lung Carcinoma (NSCLC, C=2) subtyping with 5-fold cross-validation (CV) on TCGA (n=1,041), with external validation on CPTAC (n=1,091) and NLST (n=1,008) [39]; ISUP grading from the PANDA prostate cancer challenge (C=6, n=10,616) [40]; and BRACS breast carcinoma subtyping at coarse (BRACS-C, C=3) and fine (BRACS-F, C=7) granularity (n=547) [41]. We report balanced accuracy for the multiclass tasks, AUROC for NSCLC, and weighted κ for PANDA.

Molecular biomarker prediction. We evaluate on 13 molecular biomarker prediction tasks across four cohorts. On TCGA GBMLGG, we perform 5-fold CV for IDH1 mutation prediction (GBMLGG-C, C=2, n=1,539) and histomolecular subtyping (GBMLGG-F, C=5, n=1,516), with external evaluation on the EBRAINS glioma cohort (n=894) [38]. 5-fold CV on TCGA Lung (LUAD + LUSC) for TP53, KRAS, STK11, and EGFR mutation status (C=2, n=1,024) [42]; 5-fold CV on TCGA BRCA for receptor status (ER, PR, HER2; C=2) and PIK3CA mutation (C=2), with n ranging from 741 (HER2) to 1,061 (PIK3CA); and 10-fold CV on BCNB (n=1,058) [43] for ER, PR, and HER2 status. We report AUROC for binary tasks, balanced accuracy for GBMLGG-F, and macro-averaged performance across cohorts where external evaluation is available.

Survival prediction. We predict disease-specific survival (DSS) with 5-fold CV on four TCGA cohorts [42]: breast carcinoma (BRCA, n=1,041), colon adenocarcinoma (COAD, n=417), lung adenocarcinoma (LUAD, n=456), and lung squamous cell carcinoma (LUSC, n=471). LUAD, LUSC, and COAD include external validation on CPTAC; LUAD and LUSC additionally include NLST. Performance is measured by the concordance index (C-index).

Prototype source data. Prototypes are built from HEST-1k [17], which shares no patients or slides with any downstream cohort. One HEST slide per organ is held out for interpretability (Section 5.2).

4.2 Baselines

We compare MIST against the standard linear layer across eight MIL aggregators: ABMIL [24], CLAM [13], DSMIL [25], TransMIL [12], ILRA [14], Transformer [44, 45], MeanMIL, and MaxMIL. Each aggregator retains its published hyperparameters. We further combine MIST with MAMMOTH [6], with an H&E-only prototype control, to examine whether modality bottlenecks and morphology-focused capacity bottlenecks are complementary. Further details are provided in Appendix E.

4.3 Implementation Details

All slides are processed at 20× magnification and tiled into non-overlapping 256×256 patches. Patch features are extracted using the frozen UNI2-h pathology foundation model [10]. Molecular prototypes are constructed once per organ from paired ST and H&E data following Section 3.1. Unless otherwise specified, all models are trained using AdamW with cosine learning rate decay and evaluated under the same optimization setting across all MIL aggregators. Survival, biomarker, and NSCLC tasks use matched cross-validation splits for Base and MIST; PANDA, BRACS, and EBRAINS use predefined splits with 1,000 bootstrap resamples of the test set. Additional implementation details and hyperparameters are provided in the appendix D.

Table 1: **Survival prediction.** Concordance index across four datasets using eight MIL aggregators. LUAD and LUSC are averaged across TCGA, CPTAC, and NLST, while COAD is averaged across TCGA and CPTAC. Results use matched 5-fold CV splits for Base and MIST; standard errors across folds are in parentheses. Trans.,Transformer.

Cohort	Status	ABMIL	CLAM	TransMIL	Trans.	ILRA	Mean	Max	DSMIL	Avg.
BRCA	Base	$60.58_{(6.44)}$	$60.79_{(6.33)}$	$62.66_{(9.88)}$	$63.12_{(8.92)}$	$57.47_{(6.20)}$	$65.77_{(7.26)}$	$54.95_{(3.82)}$	$60.99_{(4.92)}$	$60.79_{(3.14)}$
	+Ours	$69.34_{(3.87)}$	$69.70_{(3.65)}$	$67.56_{(7.98)}$	$70.28_{(8.02)}$	$65.31_{(6.37)}$	$70.96_{(1.54)}$	$58.77_{(9.11)}$	$68.12_{(2.51)}$	$67.50_{(3.70)}$
	Δ	+8.76	+8.91	+4.90	+7.16	+7.84	+5.19	+3.81	+7.13	+6.71
LUAD	Base	$66.07_{(6.56)}$	$65.75_{(6.72)}$	$60.90_{(5.23)}$	$56.15_{(4.79)}$	$51.03_{(3.37)}$	$66.62_{(6.40)}$	$49.40_{(4.84)}$	$64.81_{(5.63)}$	$60.09_{(6.57)}$
	+Ours	$67.49_{(6.54)}$	$67.33_{(6.85)}$	$65.64_{(9.18)}$	$62.32_{(5.90)}$	$55.94_{(3.62)}$	$67.64_{(6.89)}$	$58.74_{(4.41)}$	$66.37_{(6.32)}$	$63.93_{(4.19)}$
	Δ	+1.42	+1.59	+4.74	+6.17	+4.91	+1.02	+9.35	+1.55	+3.84
LUSC	Base	$61.99_{(2.75)}$	$61.27_{(1.81)}$	$53.79_{(4.86)}$	$54.71_{(3.01)}$	$49.29_{(6.68)}$	$61.79_{(2.65)}$	$47.14_{(4.98)}$	$57.87_{(2.58)}$	$55.98_{(5.37)}$
	+Ours	$63.41_{(2.88)}$	$63.48_{(2.11)}$	$62.54_{(4.24)}$	$61.53_{(5.50)}$	$58.72_{(4.03)}$	$62.55_{(3.26)}$	$54.30_{(4.15)}$	$61.48_{(5.04)}$	$61.00_{(2.90)}$
	Δ	+1.42	+2.22	+8.75	+6.82	+9.43	+0.75	+7.16	+3.61	+5.02
COAD	Base	$60.71_{(6.18)}$	$60.76_{(6.86)}$	$59.62_{(8.47)}$	$55.99_{(4.25)}$	$48.34_{(3.31)}$	$60.44_{(7.28)}$	$49.97_{(8.26)}$	$57.61_{(8.23)}$	$56.68_{(4.63)}$
	+Ours	$62.48_{(5.70)}$	$62.68_{(4.90)}$	$64.53_{(6.37)}$	$60.49_{(2.59)}$	$58.75_{(2.68)}$	$61.49_{(6.47)}$	$58.28_{(1.89)}$	$65.23_{(2.27)}$	$61.74_{(2.34)}$
	Δ	+1.77	+1.91	+4.91	+4.49	+10.41	+1.05	+8.31	+7.61	+5.06

Table 2: **Molecular biomarker prediction.** Mean performance across biomarker tasks within each dataset, with GBMLGG additionally averaged across TCGA and EBRAINS. Results use matched 5-fold CV splits for Base and MIST (10-fold for BCNB); standard errors across folds are in parentheses.

Dataset	Status	ABMIL	CLAM	TransMIL	Trans.	ILRA	Mean	Max	DSMIL	Avg.
GBMLGG (2 tasks)	Base	$77.63_{(1.72)}$	$77.80_{(1.47)}$	$74.36_{(1.69)}$	$76.05_{(1.39)}$	$70.29_{(1.83)}$	$76.03_{(1.47)}$	$73.16_{(1.71)}$	$76.25_{(2.41)}$	$75.20_{(2.35)}$
	+Ours	$79.28_{(1.73)}$	$79.30_{(1.50)}$	$77.33_{(1.83)}$	$78.12_{(1.50)}$	$73.26_{(1.62)}$	$77.30_{(1.48)}$	$75.49_{(1.35)}$	$77.91_{(1.52)}$	$77.25_{(1.89)}$
	Δ	+1.65	+1.49	+2.97	+2.06	+2.97	+1.27	+2.33	+1.66	+2.05
BCNB (3 tasks)	Base	$85.12_{(2.07)}$	$85.49_{(2.10)}$	$82.46_{(2.39)}$	$83.36_{(2.39)}$	$78.99_{(3.21)}$	$81.83_{(2.41)}$	$79.89_{(3.52)}$	$84.16_{(2.21)}$	$82.66_{(2.20)}$
	+Ours	$86.38_{(1.88)}$	$86.41_{(1.94)}$	$84.36_{(2.06)}$	$86.02_{(1.88)}$	$82.03_{(2.63)}$	$83.31_{(2.06)}$	$83.21_{(2.06)}$	$84.86_{(2.07)}$	$84.57_{(1.53)}$
	Δ	+1.26	+0.92	+1.89	+2.67	+3.04	+1.48	+3.32	+0.69	+1.91
BRCA (4 tasks)	Base	$76.35_{(2.38)}$	$76.36_{(2.90)}$	$71.96_{(2.63)}$	$72.22_{(2.61)}$	$68.11_{(1.88)}$	$73.33_{(2.32)}$	$71.16_{(2.70)}$	$73.70_{(2.40)}$	$72.90_{(2.55)}$
	+Ours	$77.16_{(1.90)}$	$77.21_{(1.85)}$	$75.25_{(1.94)}$	$75.64_{(1.90)}$	$71.42_{(2.15)}$	$74.32_{(2.03)}$	$74.11_{(2.03)}$	$74.83_{(2.25)}$	$74.99_{(1.73)}$
	Δ	+0.81	+0.84	+3.29	+3.42	+3.32	+0.99	+2.96	+1.13	+2.10
Lung (4 tasks)	Base	$77.79_{(2.48)}$	$77.67_{(2.32)}$	$74.73_{(2.00)}$	$74.81_{(1.92)}$	$72.06_{(1.53)}$	$76.18_{(1.85)}$	$69.83_{(3.05)}$	$75.01_{(2.86)}$	$74.76_{(2.53)}$
	+Ours	$79.29_{(1.83)}$	$79.44_{(1.60)}$	$77.68_{(2.18)}$	$77.97_{(2.54)}$	$75.26_{(1.83)}$	$77.16_{(1.95)}$	$73.00_{(2.18)}$	$77.24_{(2.11)}$	$77.13_{(1.99)}$
	Δ	+1.51	+1.77	+2.95	+3.16	+3.20	+0.98	+3.18	+2.22	+2.37

5 Results

5.1 Downstream Task Performance

Across all 14 summary results in Tables 1–3 (4 survival cohorts, 4 biomarker datasets, and 6 subtyping tasks), MIST improves the aggregator-averaged performance in all cases (mean gain $+3.5$ points, 95% CI $[2.6, 4.4]$; one-sided Wilcoxon signed-rank test, $p < 10^{-4}$).

Survival prediction. Table 1 shows that MIST improves mean concordance across all four datasets, yielding an average gain of $+5.2$ points. Table 7 further shows performance improvements in 71 out of 72 experimental configurations. The largest gains are observed for weaker aggregators that rely on coarse pooling operations. For example, MaxMIL achieves a $+7.2$-point improvement, approximately twice the typical gain obtained by stronger backbones such as ABMIL or CLAM. This trend is consistent with the hypothesis that survival outcomes are primarily driven by molecular state and microenvironmental composition, which H&E morphology captures only indirectly [3]. Consequently, weaker aggregators have limited capacity to fully exploit morphological information alone, causing them to benefit more substantially from molecularly-informed priors.

Molecular biomarker prediction. MIST improves mean performance on every dataset, yielding an average gain of $+2.1$ points (Table 2), with improvements observed in 110 out of 120 configurations (Table 8). MaxMIL alone achieves a $+3.0$-point gain. Because biomarker labels are defined by molecular assays and are only indirectly reflected in H&E morphology [46, 47], the molecularly informed organization provided by MIST is particularly relevant for these tasks.

Tissue subtyping. MIST improves average performance by $+3.3$ points across six tissue subtyping tasks from four datasets (Table 3), with gains observed in 59 out of 64 configurations (Table 9); MaxMIL alone improves by $+4.5$ points. The largest gains occur on BRACS, where balanced accuracy increases by $+6.39$ points on BRACS-C and $+3.37$ points on BRACS-F. Unlike NSCLC

Table 3: **Tissue subtyping.** Mean performance across tissue subtyping tasks using eight MIL aggregators. NSCLC is averaged across TCGA, CPTAC, and NLST. NSCLC uses 5-fold CV; PANDA, BRACS, and EBRAINS use predefined splits. Standard errors (in parentheses) are from 1,000 bootstrap resamples of the test set.

Task	Status	ABMIL	CLAM	TransMIL	Trans.	ILRA	Mean	Max	DSMIL	Avg.
NSCLC (AUROC)	Base	$93.92_{(0.75)}$	$94.46_{(0.74)}$	$92.19_{(1.32)}$	$92.99_{(1.38)}$	$89.09_{(1.44)}$	$90.33_{(0.82)}$	$93.70_{(0.80)}$	$93.43_{(0.93)}$	$92.51_{(1.76)}$
	+Ours	$94.55_{(0.73)}$	$94.56_{(0.71)}$	$93.40_{(1.05)}$	$93.92_{(1.09)}$	$91.27_{(1.32)}$	$91.48_{(0.93)}$	$94.30_{(0.71)}$	$94.41_{(0.74)}$	$93.49_{(1.27)}$
	Δ	+0.63	+0.10	+1.21	+0.93	+2.19	+1.15	+0.60	+0.98	+0.97
PANDA (Weighted κ)	Base	$93.40_{(0.65)}$	$93.72_{(0.60)}$	$93.42_{(0.62)}$	$92.70_{(0.77)}$	$92.94_{(0.76)}$	$87.74_{(1.18)}$	$83.05_{(1.64)}$	$94.24_{(0.60)}$	$91.40_{(3.69)}$
	+Ours	$94.15_{(0.69)}$	$94.15_{(0.59)}$	$93.92_{(0.63)}$	$94.19_{(0.60)}$	$95.47_{(0.48)}$	$92.17_{(0.85)}$	$86.29_{(1.32)}$	$93.99_{(0.73)}$	$93.04_{(2.68)}$
	Δ	+0.75	+0.43	+0.50	+1.48	+2.53	+4.43	+3.25	-0.24	+1.64
BRACS-C (Bal. acc.)	Base	$62.59_{(5.18)}$	$62.41_{(5.13)}$	$65.99_{(4.21)}$	$61.82_{(4.71)}$	$44.70_{(5.34)}$	$54.89_{(5.38)}$	$60.28_{(5.02)}$	$57.61_{(5.19)}$	$58.79_{(6.18)}$
	+Ours	$72.78_{(4.79)}$	$70.29_{(4.90)}$	$70.70_{(4.97)}$	$72.60_{(4.67)}$	$51.40_{(5.14)}$	$58.42_{(5.49)}$	$65.31_{(5.48)}$	$59.87_{(5.14)}$	$65.17_{(7.36)}$
	Δ	+10.19	+7.88	+4.71	+10.78	+6.70	+3.53	+5.03	+2.26	+6.39
BRACS-F (Bal. acc.)	Base	$36.66_{(4.70)}$	$44.02_{(4.60)}$	$38.98_{(4.79)}$	$41.15_{(5.17)}$	$29.93_{(4.60)}$	$31.14_{(4.60)}$	$28.41_{(4.24)}$	$34.59_{(5.19)}$	$35.61_{(5.23)}$
	+Ours	$44.35_{(4.97)}$	$45.22_{(5.14)}$	$45.58_{(5.22)}$	$41.90_{(4.58)}$	$30.25_{(4.24)}$	$32.20_{(4.82)}$	$36.02_{(4.52)}$	$36.36_{(4.91)}$	$38.98_{(5.68)}$
	Δ	+7.69	+1.20	+6.60	+0.75	+0.31	+1.06	+7.61	+1.77	+3.37
EBRAINS-C (Bal. acc.)	Base	$90.77_{(1.92)}$	$89.86_{(1.97)}$	$87.97_{(2.02)}$	$88.11_{(2.04)}$	$84.74_{(2.39)}$	$86.97_{(2.21)}$	$75.91_{(2.35)}$	$83.24_{(2.36)}$	$85.95_{(4.44)}$
	+Ours	$91.86_{(1.84)}$	$92.15_{(1.79)}$	$90.41_{(1.94)}$	$91.74_{(1.68)}$	$86.17_{(2.25)}$	$90.36_{(2.02)}$	$83.02_{(2.26)}$	$90.34_{(2.04)}$	$89.51_{(3.02)}$
	Δ	+1.08	+2.28	+2.45	+3.64	+1.43	+3.39	+7.12	+7.10	+3.56
EBRAINS-F (Bal. acc.)	Base	$69.36_{(2.20)}$	$71.04_{(2.19)}$	$72.76_{(2.00)}$	$72.10_{(1.96)}$	$63.46_{(2.21)}$	$69.08_{(2.20)}$	$60.23_{(2.28)}$	$64.25_{(2.20)}$	$67.79_{(4.28)}$
	+Ours	$74.87_{(2.07)}$	$74.15_{(1.97)}$	$73.82_{(2.08)}$	$75.38_{(1.97)}$	$68.12_{(2.17)}$	$73.79_{(2.09)}$	$63.78_{(2.17)}$	$68.36_{(2.18)}$	$71.53_{(3.95)}$
	Δ	+5.51	+3.11	+1.06	+3.28	+4.66	+4.70	+3.54	+4.11	+3.75

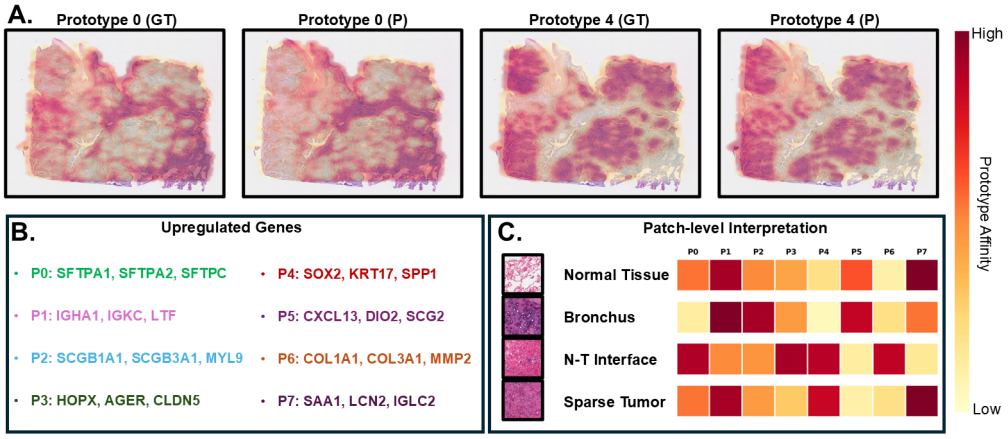


Figure 3: **Interpretability analysis of MIST.** (A) Prototype affinity maps on a held-out HEST slide with paired ST. GT and P denote transcriptomics ground truth and H&E-based prediction, respectively. (B) Top upregulated genes associated with representative molecular prototypes. (C) Patch-level affinity distributions across representative tissue regions annotated by board-certified pathologists, including normal tissue, bronchus, tumor-normal (N–T) interface, and sparse tumor.

or PANDA, several BRACS categories exhibit substantial molecular and biological overlap despite similar histological appearance, making them difficult to separate using morphology alone. In such settings, the molecularly informed organization provided by MIST helps emphasize subtle morphological cues that pathology foundation models alone do not prioritize. The gains are comparatively smaller on saturated cohorts where baseline H&E performance already exceeds 90% AUROC or weighted κ (NSCLC and PANDA), suggesting limited remaining headroom for improvement. Further details are provided in Appendix C. With CONCH [11] as the encoder, MIST also improves BRCA survival (62.50→65.99), BRCA biomarkers (72.15→73.77), and NSCLC (97.52→97.71) on average over the eight aggregators (Appendix H).

5.2 Interpretability

Spatial validation. We first evaluate whether prototypes capture spatially coherent molecular organization from H&E alone. On a held-out HEST slide with paired ST measurements, Fig. 3A compares the ground-truth spatial transcriptomic prototype maps (GT) with the corresponding model-predicted affinity maps (P) for two representative prototypes. The strong spatial agreement between GT and P indicates that MIST can infer biologically meaningful molecular distributions directly

from morphology, despite receiving no ST input at inference. Notably, the predicted patterns remain spatially structured rather than diffuse, suggesting that the prototype space captures stable tissue-level molecular organization. Further details are provided in Appendix F.

Biological grounding. To characterize the biological meaning of each prototype, Fig. 3B summarizes the top upregulated genes associated with representative molecular clusters. Several prototypes align with recognizable tissue programs and microenvironmental states. For example, P0 is enriched for surfactant-related genes (*SFTPA1*, *SFTPA2*, *SFTPC*), consistent with alveolar epithelium, while P4 captures squamous and basal-like tumor states through markers including *SOX2* and *KRT17*. These two prototypes correspond closely to distinct normal and tumor-associated tissue programs visible in the slide-level and patch-level analyses. Importantly, these biological structures emerge without explicit tissue annotations during downstream training, suggesting that the prototype space is organized around intrinsic tissue biology rather than the downstream task objective alone.

Pathological validation. We further examine whether the prototype affinities correspond to interpretable histopathological patterns at the local tissue level. Fig. 3C shows the prototype affinity profiles of four representative patches annotated by board-certified pathologists. Each tissue state exhibits a distinct prototype signature. Normal Tissue preferentially activates alveolar-associated prototypes, whereas Sparse Tumor shows elevated tumor-associated activity. Bronchus activates airway and immune-related prototypes, while the N–T Interface co-activates alveolar, tumor, and stromal-associated prototypes, reflecting the heterogeneous composition commonly observed at invasive tumor boundaries. Importantly, these prototype signatures are not directly supervised by pathology annotations, but emerge indirectly from ST-derived molecular priors, suggesting that MIST captures clinically meaningful tissue heterogeneity through the alignment between morphology and molecular organization.

5.3 Ablation and Hyperparameter Analysis

Model design ablations. Table 4 evaluates four key design choices in MIST. Replacing gene-derived prototypes with random vectors produces the largest drop (-2.2), indicating that prototype structure matters. Replacing cross-modal prototypes with H&E-only clustering, with identical architecture and capacity, still reduces performance (-1.1), suggesting that morphology alone captures only part of the underlying structure and that cross-modal anchoring remains important. Removing the low-rank factorization (-1.3) or assigning independent $\mathbf{V}_k$ to each prototype (-1.8) both reduce performance, supporting the design choice of learning prototype-specific modulation over a shared morphological subspace. Finally, removing either the view-level residual or the external residual leads to consistent degradation (-1.8), indicating that both the molecular and original pathways contribute complementary information. Using a single molecular view (72.0 on average) also underperforms the full model, supporting multi-view integration.

Key hyperparameter. Table 5 evaluates the number of prototypes per organ $K \in \{4, 8, 16, 32, 64\}$. MIST remains stable across most datasets, with only minor variation on EBRAINS-C/F, GBMLGG, and LUAD. BRACS shows greater sensitivity to prototype granularity, particularly on the coarse classification setting, suggesting that breast lesion subtypes may require finer molecular partitioning to resolve subtle morphological differences. Across all datasets, performance consistently peaks at small to moderate values of K (4–8), while larger prototype sets provide little additional benefit. This behavior suggests that only a compact set of molecular concepts is needed to effectively organize the morphology space accessible from H&E images. Further analysis of the computational cost and parameter overhead of MIST is provided in Appendix G.

Modality bottleneck. Table 6 examines whether the gains from MIST overlap with capacity-focused MIL redesigns on LUAD. Adding MAMMOTH to ABMIL improves c-index by $+0.0213$, while MIST alone yields a larger gain of $+0.0358$. Applying MIST on top of ABMIL+MAMMOTH produces an additional $+0.0323$ improvement, which remains close to the gain obtained on vanilla ABMIL. In contrast, an H&E-only variant on top of MAMMOTH does not improve it (0.6272; Appendix I), indicating that the gain is not explained by additional prototype-based capacity alone. Since both methods reorganize features at the projection stage, their gains are not expected to be strictly additive.

Table 4: Ablation of key design choices in MIST, including prototype source, view transform, number of views, and residual structure. Avg. denotes the mean performance across GBMLGG, LUAD, BRACS-C/F, and EBRAINS-C/F.

Variant	**Avg.**
MIST (full)	**73.9**
Prototype source (Eq. 1)	
Random vectors	71.7 (-2.2)
H&E-only clustering	72.8 (-1.1)
View transform (Eq. 3)	
Full-rank $\mathbf{W}_k$	72.6 (-1.3)
Unshared $\mathbf{V}_k$ per prototype	72.1 (-1.8)
Number of views	
Single view (avg.)	72.0 (-1.9)
Residual structure	
No external residual (Eq. 4)	72.1 (-1.8)
No view residual (Eq. 3)	72.1 (-1.8)

Table 5: Sensitivity analysis over the number of molecular prototypes K.

Dataset	k=4	k=8	k=16	k=32	k=64
BRACS-C	68.66	72.78	70.92	67.84	68.66
BRACS-F	44.45	44.35	43.45	42.97	44.11
EBRAINS-C	91.86	91.57	91.71	91.84	91.58
EBRAINS-F	74.87	74.37	74.29	74.45	74.01
GBMLGG	79.28	79.26	79.23	79.27	78.89
LUAD	67.47	67.49	67.24	67.42	66.94

Table 6: Performance of MIST with ABMIL and ABMIL+MAMMOTH on LUAD. Results are reported as c-index.

	w/o MIST	w/ MIST
ABMIL	0.6111	0.6496
ABMIL + MAMMOTH	0.6324	0.6647

6 Conclusion

We introduced MIST, a plug-and-play replacement for the MIL projection layer that injects ST priors into frozen pathology foundation model pipelines while requiring only H&E images at inference. Unlike prior ST-based approaches that adapt the encoder, MIST performs cross-modal adaptation at the projection layer, so that adopting a new pathology foundation model only requires rebuilding the prototype bank. By constructing cross-modal prototypes from paired ST and H&E data, MIST reorganizes patch representations according to molecular programs. Across 23 downstream tasks and 8 MIL aggregators, MIST consistently improves performance, particularly on survival prediction. Gains are strongest for weaker aggregators and molecularly heterogeneous tasks, consistent with the view that the MIL projection layer is limited not only by capacity but also by the lack of molecular supervision. Interpretability analyses further show that learned prototype affinities capture spatially coherent tissue programs directly from H&E images.

Acknowledgments and Disclosure of Funding

This research is supported by the National Medical Research Council (NMRC) Healthy and Meaningful Longevity – Cognition Grant (NICCOG2024-0028) and the AI for Public Health Program in Saw Swee Hock School of Public Health, National University of Singapore. This work was also supported by the Dame Julia Higgins Postdoc Collaborative Research Fund, the NUS-Guangzhou Research Translation and Innovation Institute Scholarship, the National Natural Science Foundation of China (Grant No. 62602189), and the China Postdoctoral Science Foundation (Grant Nos. 2026M791622, 2026T019HN, and GZB20260396). The authors declare no competing interests.

References

[1] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, 2018.

[2] Jakob Nikolas Kather, Alexander T Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*, 25(7):1054–1056, 2019.

[3] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer cell*, 40(8):865–878, 2022.

[4] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[5] Katherine A Hoadley, Christina Yau, Toshinori Hinoue, Denise M Wolf, Alexander J Lazar, Esther Drill, Ronglai Shen, Alison M Taylor, Andrew D Cherniack, Vésteinn Thorsson, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304, 2018.

[6] Daniel Shao, Joel Runevic, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, Andrew H. Song, and Faisal Mahmood. Mixture of mini experts: Overcoming the linear layer bottleneck in multiple instance learning. In *International Conference on Learning Representations (ICLR)*, 2026.

[7] Konstantin Hemker, Andrew H. Song, Cristina Almagro-Perez, Guillaume Jaume, Sophie J. Wagner, Anurag Vaidya, Nikola Simidjievski, Mateja Jamnik, and Faisal Mahmood. Towards spatial transcriptomics-driven pathology foundation models. *arXiv preprint arXiv:2602.14177*, 2026.

[8] Minsoo Lee, Jonghyun Kim, Juseung Yun, Sunwoo Yu, and Jongseong Jang. Mint: Molecularly informed training with spatial transcriptomics supervision for pathology foundation models. *arXiv preprint arXiv:2603.07895*, 2026.

[9] Ekaterina Redekop, Mara Pleasure, Zichen Wang, Kimberly Flores, Anthony Sisk, William Speier, and Corey W Arnold. Spade: Spatial transcriptomics and pathology alignment using a mixture of data experts for an expressive latent space. *arXiv preprint arXiv:2506.21857*, 2025.

[10] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

[11] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

[12] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

[13] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[14] Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *The Eleventh International Conference on Learning Representations*, 2023.

[15] Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11343–11352, 2024.

[16] Shino Magaki, Seyed A Hojat, Bowen Wei, Alexandra So, and William H Yong. An introduction to the performance of immunohistochemistry. In *Biobanking: methods and protocols*, pages 289–298. Springer, 2018.

[17] Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Perez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems*, December 2024.

[18] Jinxi Xiang, Siyu Hou, Yuchen Li, Ryan Quinton, Xiaoming Zhang, Feyisope Eweje, Xiangde Luo, Yijiang Chen, Zhe Li, Colin Bergstrom, et al. A multimodal foundation model of spatial transcriptomics and histology for biological discovery and clinical prediction. *arXiv preprint arXiv:2604.03630*, 2026.

[19] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, 30(10):2924–2935, 2024.

[20] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

[21] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16144–16155, 2022.

[22] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025.

[23] Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature medicine*, pages 1–13, 2025.

[24] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[25] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.

[26] Zhengrui Guo, Qichen Sun, Jiabo Ma, Lishuang Feng, Jinzhuo Wang, and Hao Chen. Context matters: Query-aware dynamic long sequence modeling of gigapixel images. *arXiv preprint arXiv:2501.18984*, 2025.

[27] Yucheng Xing, Hailan Mo, Zi Wang, Ling Huang, and Mengling Feng. Evidential fusion network for multimodal survival prediction under missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 238–248, 2026.

[28] Ling Huang, Yucheng Xing, Thierry Denoeux, and Mengling Feng. An evidential time-to-event prediction model based on gaussian random fuzzy numbers. In *International Conference on Belief Functions*, pages 49–57. Springer, 2024.

[29] Ling Huang, Yucheng Xing, Swapnil Mishra, Thierry Denœux, and Mengling Feng. Evidential time-to-event prediction with calibrated uncertainty quantification. *International Journal of Approximate Reasoning*, 181:109403, 2025.

[30] Ling Huang, Yucheng Xing, Qika Lin, Jinming Duan, Su Ruan, and Mengling Feng. Esurvfusion: An evidential multimodal survival fusion model based on epistemic random fuzzy sets. *IEEE Transactions on Fuzzy Systems*, 2025.

[31] Yucheng Xing, Ling Huang, Pei Liu, Jingying Ma, Jiaxing Xu, Kai He, and Mengling Feng. Semantic-anchored evidential fusion for domain-robust whole-slide survival analysis. In *European Conference on Computer Vision (ECCV)*, pages 150–169, 2026.

[32] Jingying Ma, Feng Wu, Yucheng Xing, Qika Lin, Tianyu Liu, Chenyu Liu, Ziyu Jia, and Mengling Feng. SCOPE: Structured prototype-guided adaptation for EEG foundation models with limited labels. *arXiv preprint arXiv:2602.17251*, 2026.

[33] Andrew H Song, Richard J Chen, Tong Ding, Drew FK Williamson, Guillaume Jaume, and Faisal Mahmood. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[34] Quoc Dang Vu, Kashif Rajpoot, Shan E Ahmed Raza, and Nasir Rajpoot. Handcrafted histological transformer (h2t): Unsupervised representation of whole slide images. *Medical image analysis*, 85:102743, 2023.

[35] Andrew H Song, Richard J Chen, Guillaume Jaume, Anurag J Vaidya, Alexander S Baras, and Faisal Mahmood. Multimodal prototyping for cancer survival prediction. *arXiv preprint arXiv:2407.00224*, 2024.

[36] Yucheng Xing, Ling Huang, Jingying Ma, Ruping Hong, Jiangdong Qiu, Pei Liu, Kai He, Huazhu Fu, and Mengling Feng. DPsurv: Dual-prototype evidential fusion for uncertainty-aware and interpretable whole slide image survival prediction. In *International Conference on Machine Learning (ICML)*, pages 140060–140084, 2026.

[37] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902, 2019.

[38] Thomas Roetzer-Pejrimovsky, Anna-Christina Moser, Baran Atli, Clemens Christian Vogel, Petra A Mercea, Romana Prihoda, Ellen Gelpi, Christine Haberler, Romana Höftberger, Johannes A Hainfellner, et al. The digital brain tumour atlas, an open histopathology resource. *Scientific Data*, 9(1):55, 2022.

[39] Joshua D Campbell, Anton Alexandrov, Jaegil Kim, Jeremiah Wala, Alice H Berger, Chandra Sekhar Pedamallu, Sachet A Shukla, Guangwu Guo, Angela N Brooks, Bradley A Murray, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature genetics*, 48(6):607–616, 2016.

[40] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester Van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022.

[41] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022:baac093, 2022.

[42] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[43] Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in oncology*, 11:759007, 2021.

[44] Sophia J Wagner, Daniel Reisenbüchler, Nicholas P West, Jan Moritz Niehues, Jiefu Zhu, Sebastian Foersch, Gregory Patrick Veldhuizen, Philip Quirke, Heike I Grabsch, Piet A van den Brandt, et al. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer cell*, 41(9):1650–1661, 2023.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[46] Jakob Nikolas Kather, Lara R Heij, Heike I Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M Niehues, Kai AJ Sommer, Peter Bankhead, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature cancer*, 1(8):789–799, 2020.

[47] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature cancer*, 1(8):800–810, 2020.

Appendix

A Molecular Prototype Construction

We derive K molecular prototypes offline, each capturing a distinct gene expression program while living in the morphological feature space so it can directly interact with patch embeddings at training time. We use the HEST-1k dataset [17], which provides 1,276 spatially-resolved transcriptomic profiles paired with H&E whole-slide images across 26 organ types and approximately 940,000 human Visium spots.

Data curation. We apply SEAL-style filtering [7] to select high-quality samples: (1)Homo sapiens only; (2) Visium or Visium HD (spot diameter $\leq 55\,\mu$m), excluding legacy $100\,\mu$m ST arrays; (3) 14 organ types (Breast, Bowel, Skin, Kidney, Heart, Prostate, Brain, Lung, Uterus, Liver, Bladder, Pancreas, Cervix, Ovary); (4) $\geq 5{,}000$ detected genes per sample.

Patch features. For each sample, patches at $0.5\,\mu$m/px ($20\times$) are tessellated from the H&E slide and encoded by a frozen pathology foundation model (UNI2-h [10], D=1536). Each patch is matched to its overlapping Visium spot by coordinate alignment, producing paired samples $\{(\mathbf{x}_j, \mathbf{g}_j)\}_{j=1}^{M}$.

Gene expression preprocessing. After quality control (removing empty spots and genes detected in $<10\%$ of spots), counts are normalized to a target sum of 10^4 per spot and transformed via $\log(1+x)$. We select the top 2,000 highly variable genes (HVGs) using the Seurat v3 variance-stabilizing method [37] on the raw count layer. To mitigate Visium sparsity and noise, we apply spatial smoothing within each tissue section independently to prevent leakage across sections. For each spot j, we average its HVG vector with its n=6 nearest neighbors by Euclidean distance on spatial coordinates (approximating the Visium hexagonal grid):

$$\bar{\mathbf{g}}_j = \frac{1}{|\mathcal{N}_n(j)| + 1}\Big(\mathbf{g}_j + \sum_{j' \in \mathcal{N}_n(j)} \mathbf{g}_{j'}\Big). \tag{6}$$

Molecular clustering. We run MiniBatch K-means on the smoothed vectors $\{\bar{\mathbf{g}}_j\}_{j=1}^{M}$ pooled across all samples, assigning each spot a cluster label $c_j \in \{1, \ldots, K\}$. Each cluster corresponds to a distinct molecular program, e.g., hepatocyte metabolism (SERPINA1$^+$), B-cell immune response (IGKC$^+$), glial identity (GFAP$^+$), or epithelial differentiation (KRT19$^+$). For each organ, we scan $K \in \{4, 8, 16, 32, 64\}$ and select the value with the highest silhouette score on a random subset of 50,000 spots (e.g., K=8 for breast and K=4 for brain).

Prototype computation. For each cluster k, the molecular prototype is the centroid of its patch features):

$$\boldsymbol{\mu}_k = \frac{1}{|\mathcal{C}_k|} \sum_{j \in \mathcal{C}_k} \mathbf{x}_j \in \mathbb{R}^D, \qquad \mathcal{C}_k = \{j : c_j = k\}. \tag{7}$$

This cross-modal design is central: clustering is performed in gene expression space to discover molecular states, but the prototypes live in morphological feature space so they can directly condition patch embeddings at training time. Each $\boldsymbol{\mu}_k$ thus encodes the average morphological appearance of a specific molecular program, serving as a *molecular view*. The full set $\boldsymbol{\Pi} = \{\boldsymbol{\mu}_k\}_{k=1}^{K} \in \mathbb{R}^{K \times D}$ is stored offline and used to initialize MIST at the start of training.

B MIST Can Increase Separation Between Morphologically Similar but Molecularly Distinct Patches

Setup. Let $\mathbf{h}_i, \mathbf{h}_j \in \mathbb{R}^d$ be two patch embeddings that are morphologically similar but molecularly distinct:

$$\mathbf{h}_i - \mathbf{h}_j = \epsilon \mathbf{v}, \quad \|\mathbf{v}\| = 1, \quad \epsilon \ll 1, \tag{8}$$

and patch i is associated with prototype A while patch j is associated with prototype B, i.e., $\alpha_A(\mathbf{h}_i) > \alpha_A(\mathbf{h}_j)$. Denote $\Delta\alpha_k = \alpha_k(\mathbf{h}_i) - \alpha_k(\mathbf{h}_j)$, with $\Delta\alpha_A > 0$ and $\Delta\alpha_B < 0$.

Molecular baseline. Define the *molecular baseline* as the MIST output when the affinity is uniform across prototypes, i.e., $\alpha_k \equiv 1/K$ for all k. In this case the prototype-conditioned term reduces to a fixed transform $\bar{\mathbf{W}} = \frac{1}{K}\sum_k \mathbf{U}_k\mathbf{V}$, and the output distance is

$$\|f_{\text{base}}(\mathbf{h}_i) - f_{\text{base}}(\mathbf{h}_j)\| = \|(2\mathbf{I} + \bar{\mathbf{W}})\epsilon\mathbf{v}\|. \tag{9}$$

This baseline captures the separation that MIST achieves from morphology alone, with no molecular routing.

MIST output distance. The full MIST output difference (to first order in ϵ) is

$$\tilde{\mathbf{h}}_i - \tilde{\mathbf{h}}_j = (2\mathbf{I} + \bar{\mathbf{W}})\epsilon\mathbf{v} + \underbrace{\Big(\sum_k \Delta\alpha_k\,\mathbf{U}_k\Big)\mathbf{V}\bar{\mathbf{h}}}_{\text{molecular routing term}}, \tag{10}$$

where $\bar{\mathbf{h}} = \frac{1}{2}(\mathbf{h}_i + \mathbf{h}_j)$ and higher-order terms in ϵ are absorbed into $o(\epsilon)$.

Proposition. *There exists a parameter configuration* $\{\mathbf{U}_k\}, \mathbf{V}$ *such that*

$$\|\tilde{\mathbf{h}}_i - \tilde{\mathbf{h}}_j\| > \|f_{\text{base}}(\mathbf{h}_i) - f_{\text{base}}(\mathbf{h}_j)\|. \tag{11}$$

Proof. Let $\mathbf{n} = (2\mathbf{I} + \bar{\mathbf{W}})\mathbf{v}$ denote the baseline direction (normalized: $\hat{\mathbf{n}} = \mathbf{n}/\|\mathbf{n}\|$). Construct parameters as follows:

$$\mathbf{V} = \hat{\mathbf{h}}^\top, \qquad \mathbf{U}_A - \mathbf{U}_B = c\,\hat{\mathbf{n}}, \quad c > 0, \tag{12}$$

where $\hat{\mathbf{h}} = \bar{\mathbf{h}}/\|\bar{\mathbf{h}}\|$. Under this construction, the molecular routing term becomes

$$\Big(\sum_k \Delta\alpha_k\,\mathbf{U}_k\Big)\mathbf{V}\bar{\mathbf{h}} = \Delta\alpha_A\,(\mathbf{U}_A - \mathbf{U}_B)\,\|\bar{\mathbf{h}}\| = c\,\Delta\alpha_A\,\|\bar{\mathbf{h}}\|\,\hat{\mathbf{n}}, \tag{13}$$

using $\sum_k \Delta\alpha_k = 0$ (from ℓ_1 normalization). Since $\Delta\alpha_A > 0$, $c > 0$, and $\|\bar{\mathbf{h}}\| > 0$, this term is strictly positive along $\hat{\mathbf{n}}$, i.e., aligned with the baseline direction. Therefore,

$$\|\tilde{\mathbf{h}}_i - \tilde{\mathbf{h}}_j\| = \|\epsilon\mathbf{n} + c\,\Delta\alpha_A\,\|\bar{\mathbf{h}}\|\,\hat{\mathbf{n}}\| = \epsilon\|\mathbf{n}\| + c\,\Delta\alpha_A\,\|\bar{\mathbf{h}}\| > \epsilon\|\mathbf{n}\| = \|f_{\text{base}}(\mathbf{h}_i) - f_{\text{base}}(\mathbf{h}_j)\|, \tag{14}$$

which establishes Eq. (11).

Remark. The construction in Eq. (12) is not unique; any $\mathbf{U}_A - \mathbf{U}_B$ with a positive projection onto $\hat{\mathbf{n}}$ suffices. The key condition is $\sum_k \Delta\alpha_k = 0$, which holds by design from the ℓ_1 normalization of affinities, and $\Delta\alpha_A > 0$, which holds whenever the prototype affinities correctly reflect the molecular identity of the patch. Since affinities are computed from H&E features, this is an assumption on the learned prototypes rather than a guarantee. Whether MIL training discovers such a configuration is an empirical question; Section 5.2 provides evidence that it does.

C Detailed Results

This section summarizes the complete downstream evaluation results across survival prediction, molecular biomarker prediction, and tissue subtyping. In total, we evaluate 256 model-task configurations spanning multiple datasets, external validation cohorts, and eight MIL aggregators. MIST improves 240 configurations, with only 16 configurations showing negative changes relative to the standard projection layer.The detailed tables below show that the gains are broadly consistent across endpoint types, aggregators, and external evaluation settings (Tables 8, 9, and 7).

D Implementation Details

All whole-slide images are processed at $20\times$ magnification ($0.5\,\mu$m/pixel) and tiled into non-overlapping 256×256 patches. Patch embeddings are extracted using the frozen UNI2-h encoder [10], a ViT-H/14 pathology foundation model pretrained with the DINOv2 objective, producing $1{,}536$-dimensional feature vectors. Molecular prototypes are constructed once per organ from paired ST

Table 7: **Survival prediction (detailed breakdown).** Concordance index (%) across individual cohorts and external validation datasets using eight MIL aggregators. CPTAC and NLST are evaluated as external cohorts. Matched 5-fold CV splits for Base and MIST; standard errors across folds are in parentheses. Avg. denotes the cross-aggregator mean performance.

Cohort	Status	ABMIL	CLAM	TransMIL	Trans.	ILRA	Mean	Max	DSMIL	Avg.
BRCA	Base	60.58$_{(6.44)}$	60.79$_{(6.33)}$	62.66$_{(9.88)}$	63.12$_{(8.92)}$	57.47$_{(6.20)}$	65.77$_{(7.26)}$	54.95$_{(3.82)}$	60.99$_{(4.92)}$	60.79$_{(3.14)}$
	+Ours	69.34$_{(3.87)}$	69.70$_{(3.65)}$	67.56$_{(7.98)}$	70.28$_{(8.02)}$	65.31$_{(6.37)}$	70.96$_{(1.54)}$	58.77$_{(9.11)}$	68.12$_{(2.51)}$	67.50$_{(3.70)}$
	Δ	+8.76	+8.91	+4.90	+7.16	+7.84	+5.19	+3.81	+7.13	+6.71
LUAD (TCGA)	Base	61.11$_{(11.26)}$	61.00$_{(11.39)}$	63.45$_{(7.33)}$	58.80$_{(6.42)}$	47.53$_{(1.26)}$	61.40$_{(10.83)}$	52.26$_{(7.68)}$	58.76$_{(8.76)}$	58.04$_{(5.05)}$
	+Ours	64.96$_{(10.77)}$	64.80$_{(10.85)}$	63.85$_{(15.56)}$	60.27$_{(9.55)}$	47.88$_{(2.83)}$	63.37$_{(11.57)}$	55.05$_{(4.89)}$	62.75$_{(9.87)}$	60.37$_{(5.61)}$
	Δ	+3.85	+3.81	+0.40	+1.48	+0.35	+1.97	+2.79	+3.99	+2.33
LUAD (CPTAC, ext.)	Base	73.59$_{(1.44)}$	74.19$_{(2.07)}$	65.54$_{(5.18)}$	53.53$_{(5.17)}$	53.73$_{(5.31)}$	73.76$_{(1.98)}$	49.76$_{(1.43)}$	74.31$_{(4.05)}$	64.80$_{(10.08)}$
	+Ours	73.69$_{(3.01)}$	74.68$_{(2.58)}$	71.70$_{(1.89)}$	66.20$_{(3.22)}$	67.61$_{(4.29)}$	74.21$_{(2.51)}$	67.07$_{(5.40)}$	74.70$_{(3.89)}$	71.23$_{(3.44)}$
	Δ	+0.10	+0.49	+6.16	+12.67	+13.88	+0.45	+17.31	+0.39	+6.43
LUAD (NLST, ext.)	Base	63.53$_{(0.61)}$	62.05$_{(1.27)}$	53.70$_{(1.24)}$	56.12$_{(0.94)}$	51.84$_{(2.09)}$	64.69$_{(1.23)}$	46.17$_{(3.03)}$	61.37$_{(1.40)}$	57.43$_{(6.13)}$
	+Ours	63.84$_{(1.77)}$	62.52$_{(4.05)}$	61.35$_{(2.66)}$	60.49$_{(1.64)}$	52.33$_{(3.60)}$	65.32$_{(1.52)}$	54.11$_{(2.31)}$	61.65$_{(2.73)}$	60.20$_{(4.29)}$
	Δ	+0.31	+0.47	+7.66	+4.38	+0.49	+0.63	+7.95	+0.28	+2.77
LUSC (TCGA)	Base	51.63$_{(4.05)}$	49.88$_{(1.87)}$	54.15$_{(7.06)}$	53.57$_{(2.32)}$	43.00$_{(5.55)}$	50.42$_{(4.45)}$	45.45$_{(6.75)}$	47.50$_{(2.90)}$	49.45$_{(3.65)}$
	+Ours	51.82$_{(3.72)}$	50.39$_{(3.20)}$	57.48$_{(6.80)}$	56.28$_{(9.06)}$	51.99$_{(5.57)}$	51.46$_{(4.62)}$	52.06$_{(1.59)}$	54.69$_{(7.37)}$	53.27$_{(2.39)}$
	Δ	+0.20	+0.51	+3.33	+2.71	+8.99	+1.04	+6.61	+7.19	+3.82
LUSC (CPTAC, ext.)	Base	65.26$_{(2.27)}$	65.42$_{(1.97)}$	48.02$_{(4.04)}$	50.96$_{(3.15)}$	51.30$_{(4.90)}$	67.96$_{(0.76)}$	46.04$_{(5.32)}$	62.48$_{(3.27)}$	57.18$_{(8.36)}$
	+Ours	68.45$_{(2.60)}$	70.78$_{(0.51)}$	67.24$_{(2.02)}$	65.94$_{(1.37)}$	53.56$_{(3.67)}$	68.05$_{(2.46)}$	60.37$_{(6.12)}$	67.71$_{(4.47)}$	65.26$_{(5.24)}$
	Δ	+3.19	+5.36	+19.23	+14.98	+2.26	+0.09	+14.33	+5.23	+8.09
LUSC (NLST, ext.)	Base	69.08$_{(1.05)}$	68.50$_{(1.58)}$	59.19$_{(2.20)}$	59.61$_{(3.43)}$	53.57$_{(8.89)}$	67.01$_{(0.89)}$	49.94$_{(0.64)}$	63.64$_{(0.92)}$	61.32$_{(6.58)}$
	+Ours	69.96$_{(2.09)}$	69.27$_{(1.69)}$	62.90$_{(1.89)}$	62.37$_{(2.61)}$	70.60$_{(2.07)}$	68.14$_{(2.13)}$	50.48$_{(3.41)}$	62.05$_{(1.43)}$	64.47$_{(6.25)}$
	Δ	+0.88	+0.77	+3.70	+2.76	+17.03	+1.13	+0.53	-1.59	+3.15
COAD (TCGA)	Base	60.71$_{(8.60)}$	59.84$_{(9.68)}$	60.20$_{(11.62)}$	58.87$_{(5.45)}$	48.38$_{(3.52)}$	58.70$_{(10.26)}$	48.92$_{(11.55)}$	57.22$_{(11.16)}$	56.61$_{(4.70)}$
	+Ours	62.08$_{(7.98)}$	63.02$_{(6.78)}$	63.73$_{(8.92)}$	65.00$_{(2.67)}$	61.57$_{(2.79)}$	60.10$_{(9.13)}$	64.96$_{(2.16)}$	69.34$_{(2.21)}$	63.73$_{(2.64)}$
	Δ	+1.38	+3.18	+3.53	+6.13	+13.20	+1.40	+16.04	+12.12	+7.12
COAD (CPTAC, ext.)	Base	60.72$_{(1.54)}$	61.69$_{(0.61)}$	59.04$_{(2.91)}$	53.12$_{(2.54)}$	48.31$_{(3.08)}$	62.18$_{(0.92)}$	51.01$_{(1.77)}$	58.01$_{(3.31)}$	56.76$_{(4.92)}$
	+Ours	62.88$_{(1.16)}$	62.33$_{(1.40)}$	65.33$_{(1.22)}$	55.97$_{(2.50)}$	55.93$_{(2.55)}$	62.88$_{(0.67)}$	51.60$_{(1.57)}$	61.11$_{(2.32)}$	59.75$_{(4.40)}$
	Δ	+2.16	+0.64	+6.29	+2.85	+7.63	+0.70	+0.58	+3.10	+3.00

and H&E data following Section 3.1. Unless otherwise specified, MIST uses a projection dimension of d=512 and a low-rank bottleneck dimension of r=16.

All models are optimized using AdamW with learning rate 10^{-4} and weight decay 10^{-5}. Training uses a batch size of 1 with gradient accumulation over 32 iterations, together with input feature dropout of 0.1 and model dropout of 0.25. A cosine learning rate schedule with one epoch of warmup is applied throughout training.

For classification tasks, models are trained for up to 20 epochs with early stopping based on validation performance (patience 5, minimum training length 10 epochs). For survival prediction, all models are trained for 20 epochs without early stopping. Unless otherwise noted, all MIL aggregators follow their published hyperparameter settings. All experiments are conducted on NVIDIA H100 GPUs.

E Baseline MIL Aggregators

We evaluate MIST as a plug-in replacement for the standard projection layer across eight representative MIL aggregators, following the benchmark configuration of [6]. These aggregators span three broad categories: pooling-based methods, attention-based methods, and transformer-based architectures.

MeanMIL and **MaxMIL** are parameter-free pooling baselines that aggregate patch embeddings using mean pooling and max pooling, respectively. Despite their simplicity, they provide useful reference points for evaluating the contribution of learned aggregation mechanisms.

ABMIL [24] aggregates patch embeddings through gated attention, assigning each patch a learnable relevance weight before weighted pooling. Its simplicity and interpretability have made it one of the most widely used MIL baselines in computational pathology.

CLAM [13] extends ABMIL with a clustering-based auxiliary objective that encourages the attention mechanism to focus on diagnostically informative regions. This auxiliary supervision improves instance selection without requiring patch-level annotations.

TransMIL [12] introduces transformer-based aggregation with morphology-aware positional encoding to model long-range spatial dependencies among patches. By explicitly capturing inter-patch interactions, it provides richer contextual representations than independent attention pooling.

Table 8: **Molecular biomarker prediction (detailed breakdown).** Results across 15 biomarker prediction tasks spanning GBMLGG, BCNB, BRCA, and Lung cohorts using eight MIL aggregators. TCGA and EBRAINS are reported separately for GBMLGG. AUROC is used for binary classification tasks and balanced accuracy for Histomol. Matched 5-fold CV splits for Base and MIST (10-fold for BCNB); standard errors across folds are in parentheses. Avg. denotes the cross-aggregator mean performance.

Task	Status	ABMIL	CLAM	TransMIL	Trans.	ILRA	Mean	Max	DSMIL	Avg.
GBMLGG IDH1 (TCGA)	Base	$95.86_{(0.38)}$	$95.94_{(0.99)}$	$94.10_{(1.37)}$	$95.34_{(1.56)}$	$94.11_{(2.20)}$	$96.24_{(0.90)}$	$92.42_{(3.56)}$	$95.12_{(0.71)}$	$94.89_{(1.20)}$
	+Ours	$96.61_{(0.82)}$	$96.61_{(0.59)}$	$95.63_{(1.43)}$	$96.38_{(0.81)}$	$96.66_{(1.36)}$	$96.57_{(0.80)}$	$94.63_{(1.29)}$	$96.62_{(0.54)}$	$96.21_{(0.68)}$
	Δ	+0.76	+0.68	+1.53	+1.03	+2.55	+0.33	+2.21	+1.50	+1.32
GBMLGG IDH1 (EBRAINS)	Base	$93.40_{(2.23)}$	$94.23_{(1.03)}$	$89.91_{(2.15)}$	$93.09_{(1.58)}$	$88.66_{(1.95)}$	$91.68_{(1.21)}$	$90.33_{(1.12)}$	$92.39_{(2.85)}$	$91.71_{(1.80)}$
	+Ours	$94.95_{(0.81)}$	$94.87_{(1.17)}$	$93.27_{(1.81)}$	$94.81_{(1.10)}$	$91.68_{(1.86)}$	$93.29_{(0.99)}$	$92.36_{(0.85)}$	$94.82_{(0.97)}$	$93.76_{(1.21)}$
	Δ	+1.55	+0.64	+3.36	+1.73	+3.02	+1.62	+2.03	+2.43	+2.04
GBMLGG Histomol (TCGA)	Base	$66.57_{(6.03)}$	$66.37_{(5.01)}$	$63.17_{(5.89)}$	$66.74_{(3.45)}$	$56.18_{(5.86)}$	$66.57_{(4.91)}$	$60.10_{(5.61)}$	$64.92_{(8.24)}$	$63.83_{(3.61)}$
	+Ours	$68.85_{(5.92)}$	$69.01_{(5.15)}$	$66.65_{(6.12)}$	$68.32_{(3.56)}$	$58.60_{(5.91)}$	$67.86_{(4.63)}$	$61.65_{(5.01)}$	$65.94_{(5.39)}$	$65.86_{(3.54)}$
	Δ	+2.28	+2.64	+3.48	+1.58	+2.41	+1.29	+1.54	+1.02	+2.03
GBMLGG Histomol (EBRAINS)	Base	$54.69_{(2.40)}$	$54.68_{(2.72)}$	$50.24_{(5.07)}$	$49.05_{(3.75)}$	$42.19_{(2.25)}$	$49.63_{(2.86)}$	$49.79_{(1.23)}$	$52.57_{(3.41)}$	$50.35_{(3.74)}$
	+Ours	$56.71_{(3.39)}$	$56.70_{(2.84)}$	$53.77_{(3.26)}$	$52.97_{(4.63)}$	$46.09_{(1.39)}$	$51.48_{(3.47)}$	$53.32_{(1.30)}$	$54.25_{(2.59)}$	$53.16_{(3.15)}$
	Δ	+2.02	+2.02	+3.54	+3.92	+3.90	+1.85	+3.54	+1.68	+2.81
BCNB ER	Base	$92.38_{(2.18)}$	$93.07_{(1.89)}$	$90.66_{(3.18)}$	$90.98_{(3.87)}$	$87.29_{(4.57)}$	$88.93_{(2.85)}$	$75.67_{(8.21)}$	$91.84_{(2.92)}$	$88.85_{(5.28)}$
	+Ours	$93.31_{(2.34)}$	$92.80_{(2.02)}$	$92.37_{(2.90)}$	$93.13_{(1.73)}$	$90.23_{(3.02)}$	$90.50_{(2.01)}$	$91.29_{(3.17)}$	$92.81_{(2.81)}$	$92.06_{(1.13)}$
	Δ	+0.92	−0.27	+1.71	+2.16	+2.95	+1.57	+15.61	+0.98	+3.20
BCNB HER2	Base	$76.54_{(4.93)}$	$76.92_{(4.44)}$	$73.56_{(5.31)}$	$74.34_{(4.89)}$	$70.21_{(6.42)}$	$72.70_{(5.03)}$	$64.10_{(6.82)}$	$74.73_{(3.73)}$	$72.89_{(3.87)}$
	+Ours	$77.87_{(4.20)}$	$77.80_{(4.97)}$	$72.81_{(5.69)}$	$77.51_{(3.83)}$	$72.58_{(6.12)}$	$74.26_{(4.15)}$	$73.57_{(3.94)}$	$75.53_{(4.39)}$	$75.24_{(2.10)}$
	Δ	+1.33	+0.88	−0.75	+3.17	+2.37	+1.56	+9.47	+0.80	+2.35
BCNB PR	Base	$86.42_{(3.12)}$	$87.06_{(2.91)}$	$83.74_{(4.80)}$	$84.75_{(3.56)}$	$79.48_{(5.53)}$	$83.87_{(4.32)}$	$80.77_{(7.23)}$	$86.44_{(3.21)}$	$84.07_{(2.57)}$
	+Ours	$87.97_{(2.96)}$	$88.09_{(2.40)}$	$86.32_{(3.53)}$	$87.43_{(3.75)}$	$83.29_{(3.98)}$	$85.18_{(4.12)}$	$84.77_{(3.55)}$	$86.18_{(3.26)}$	$86.15_{(1.57)}$
	Δ	+1.54	+1.03	+2.59	+2.67	+3.81	+1.31	+4.00	−0.26	+2.09
BRCA ER	Base	$90.61_{(2.33)}$	$91.29_{(2.93)}$	$86.30_{(5.68)}$	$86.42_{(2.83)}$	$80.78_{(3.84)}$	$87.95_{(2.49)}$	$86.13_{(3.46)}$	$89.21_{(3.60)}$	$87.34_{(3.09)}$
	+Ours	$91.19_{(2.50)}$	$90.56_{(3.38)}$	$89.45_{(2.68)}$	$90.84_{(2.84)}$	$85.45_{(2.09)}$	$88.66_{(2.26)}$	$88.40_{(2.14)}$	$89.87_{(2.21)}$	$89.30_{(1.73)}$
	Δ	+0.58	−0.73	+3.15	+4.42	+4.67	+0.70	+2.27	+0.66	+1.97
BRCA HER2	Base	$70.43_{(7.29)}$	$70.06_{(5.50)}$	$64.92_{(4.48)}$	$63.77_{(4.87)}$	$57.67_{(3.70)}$	$66.20_{(6.81)}$	$60.12_{(4.89)}$	$62.13_{(5.79)}$	$64.41_{(4.20)}$
	+Ours	$72.20_{(4.06)}$	$68.67_{(6.03)}$	$69.35_{(4.21)}$	$71.00_{(2.23)}$	$65.40_{(6.18)}$	$68.27_{(5.69)}$	$67.45_{(4.97)}$	$68.59_{(6.87)}$	$68.87_{(1.95)}$
	Δ	+1.77	−1.40	+4.44	+7.24	+7.73	+2.07	+7.34	+6.46	+4.45
BRCA PIK3CA	Base	$64.43_{(3.80)}$	$65.20_{(5.88)}$	$57.88_{(4.37)}$	$59.11_{(5.37)}$	$57.36_{(4.74)}$	$61.70_{(4.82)}$	$58.73_{(6.49)}$	$61.71_{(6.16)}$	$60.77_{(2.78)}$
	+Ours	$64.37_{(5.11)}$	$65.94_{(4.38)}$	$63.68_{(3.89)}$	$63.19_{(4.60)}$	$60.88_{(5.00)}$	$62.71_{(3.69)}$	$62.06_{(3.88)}$	$62.63_{(4.96)}$	$63.18_{(1.43)}$
	Δ	−0.07	+0.74	+5.80	+4.08	+3.52	+1.02	+3.34	+0.92	+2.42
BRCA PR	Base	$79.91_{(4.18)}$	$80.24_{(2.81)}$	$74.70_{(5.53)}$	$74.06_{(2.54)}$	$69.77_{(3.75)}$	$77.48_{(3.36)}$	$72.67_{(2.98)}$	$78.41_{(3.96)}$	$75.91_{(3.47)}$
	+Ours	$80.89_{(2.99)}$	$80.36_{(3.01)}$	$78.52_{(4.46)}$	$77.54_{(4.82)}$	$73.96_{(2.58)}$	$77.64_{(3.86)}$	$78.54_{(4.62)}$	$78.20_{(2.03)}$	$78.21_{(1.96)}$
	Δ	+0.97	+0.12	+3.82	+3.48	+4.19	+0.16	+5.87	−0.21	+2.30
Lung EGFR	Base	$71.71_{(7.74)}$	$69.68_{(7.70)}$	$70.08_{(5.08)}$	$68.24_{(7.19)}$	$67.40_{(7.64)}$	$72.02_{(4.50)}$	$59.65_{(7.54)}$	$65.72_{(8.62)}$	$68.06_{(3.75)}$
	+Ours	$73.86_{(3.72)}$	$73.37_{(3.22)}$	$69.39_{(8.39)}$	$72.37_{(5.63)}$	$74.71_{(3.01)}$	$72.68_{(3.87)}$	$65.18_{(5.40)}$	$71.21_{(5.74)}$	$71.60_{(2.87)}$
	Δ	+2.15	+3.68	−0.69	+4.13	+7.32	+0.67	+5.53	+5.49	+3.53
Lung KRAS	Base	$82.44_{(4.21)}$	$82.71_{(3.57)}$	$79.13_{(4.72)}$	$79.53_{(1.96)}$	$72.22_{(2.62)}$	$82.38_{(3.26)}$	$78.76_{(5.59)}$	$82.79_{(3.37)}$	$80.00_{(3.35)}$
	+Ours	$83.34_{(3.13)}$	$83.60_{(3.53)}$	$82.86_{(2.66)}$	$82.52_{(4.24)}$	$79.05_{(2.45)}$	$82.21_{(3.94)}$	$81.53_{(1.62)}$	$81.28_{(5.95)}$	$82.05_{(1.36)}$
	Δ	+0.90	+0.88	+3.73	+3.00	+6.84	−0.17	+2.78	−1.51	+2.05
Lung STK11	Base	$85.82_{(2.52)}$	$85.35_{(2.74)}$	$80.42_{(2.26)}$	$77.26_{(9.01)}$	$70.48_{(3.59)}$	$80.62_{(2.44)}$	$61.72_{(9.41)}$	$81.89_{(3.70)}$	$77.94_{(7.63)}$
	+Ours	$86.62_{(3.87)}$	$87.17_{(2.98)}$	$84.95_{(5.08)}$	$83.39_{(5.90)}$	$78.46_{(5.16)}$	$81.67_{(4.29)}$	$78.40_{(6.13)}$	$83.48_{(3.79)}$	$83.02_{(3.13)}$
	Δ	+0.80	+1.82	+4.53	+6.13	+7.98	+1.05	+16.69	+1.58	+5.07
Lung TP53	Base	$71.17_{(3.79)}$	$72.92_{(2.62)}$	$67.32_{(4.09)}$	$69.37_{(4.65)}$	$63.75_{(4.36)}$	$70.08_{(3.58)}$	$61.88_{(5.61)}$	$68.54_{(4.64)}$	$68.13_{(3.47)}$
	+Ours	$73.36_{(3.86)}$	$73.62_{(3.09)}$	$71.83_{(3.72)}$	$73.59_{(4.36)}$	$68.80_{(3.42)}$	$72.08_{(3.42)}$	$66.89_{(2.55)}$	$71.49_{(3.41)}$	$71.46_{(2.27)}$
	Δ	+2.19	+0.70	+4.51	+4.22	+5.06	+2.00	+5.01	+2.95	+3.33

TransformerMIL [44, 45] applies a standard transformer encoder without tissue-specific positional encoding, serving as a simplified variant of TransMIL that isolates the contribution of transformer-based sequence modeling.

ILRA [14] formulates MIL aggregation through iterative low-rank approximation of the patch interaction matrix, constraining the representation to a compact latent subspace. This design reduces overfitting while preserving dominant patterns of inter-patch variation.

DSMIL [25] adopts a dual-stream architecture in which one branch identifies the most informative instance while the second branch propagates attention scores according to patch similarity to this anchor. The combination of instance-level and bag-level objectives improves robustness and calibration.

F More Interpretability Results

We provide additional interpretability visualizations on three held-out HEST slides from brain, breast, and lung tissues. For each slide, we visualize the whole-slide prototype affinity distributions predicted from H&E alone, together with representative molecular prototypes and their associated spatial patterns.

Table 9: **Tissue subtyping (detailed breakdown).** Results across NSCLC, PANDA, BRACS, and EBRAINS tissue subtyping tasks using eight MIL aggregators. NSCLC is evaluated separately on TCGA, CPTAC, and NLST cohorts. AUROC is reported for NSCLC, weighted κ for PANDA, and balanced accuracy for BRACS and EBRAINS. NSCLC uses 5-fold CV; PANDA, BRACS, and EBRAINS use predefined splits with standard errors from 1,000 bootstrap resamples. Avg. denotes the cross-aggregator mean performance.

Task	Status	ABMIL	CLAM	TransMIL	Trans.	ILRA	Mean	Max	DSMIL	Avg.
NSCLC (TCGA, AUROC)	Base	$98.25_{(1.05)}$	$98.31_{(1.09)}$	$98.04_{(1.20)}$	$97.89_{(1.14)}$	$96.96_{(1.75)}$	$96.95_{(1.15)}$	$96.50_{(0.73)}$	$97.70_{(1.21)}$	$97.57_{(0.64)}$
	+Ours	$98.51_{(0.98)}$	$98.53_{(0.94)}$	$98.26_{(1.04)}$	$98.31_{(1.01)}$	$97.61_{(1.35)}$	$97.53_{(1.14)}$	$97.84_{(0.77)}$	$98.20_{(0.94)}$	$98.10_{(0.36)}$
	Δ	+0.26	+0.21	+0.22	+0.42	+0.65	+0.58	+1.35	+0.51	+0.52
NSCLC (CPTAC, ext.)	Base	$97.20_{(0.42)}$	$97.50_{(0.31)}$	$93.42_{(1.84)}$	$94.19_{(1.94)}$	$92.09_{(1.42)}$	$93.95_{(0.75)}$	$96.13_{(0.65)}$	$96.71_{(0.62)}$	$95.15_{(1.87)}$
	+Ours	$96.86_{(0.53)}$	$96.69_{(0.46)}$	$95.33_{(1.15)}$	$95.54_{(1.45)}$	$94.88_{(1.18)}$	$94.96_{(0.80)}$	$95.96_{(0.56)}$	$96.11_{(0.62)}$	$95.79_{(0.70)}$
	Δ	-0.34	-0.81	+1.92	+1.35	+2.78	+1.02	-0.16	-0.60	+0.64
NSCLC (NLST, ext.)	Base	$86.32_{(0.65)}$	$87.56_{(0.60)}$	$85.12_{(0.62)}$	$86.90_{(0.80)}$	$78.21_{(1.09)}$	$80.10_{(0.39)}$	$88.47_{(0.99)}$	$85.87_{(0.86)}$	$84.82_{(3.44)}$
	+Ours	$88.29_{(0.58)}$	$88.45_{(0.63)}$	$86.60_{(0.94)}$	$87.93_{(0.65)}$	$81.34_{(1.43)}$	$81.96_{(0.83)}$	$89.09_{(0.79)}$	$88.92_{(0.63)}$	$86.57_{(2.93)}$
	Δ	+1.97	+0.89	+1.48	+1.03	+3.13	+1.86	+0.61	+3.04	+1.75
PANDA (Weighted κ)	Base	$93.40_{(0.65)}$	$93.72_{(0.60)}$	$93.42_{(0.62)}$	$92.70_{(0.77)}$	$92.94_{(0.76)}$	$87.74_{(1.18)}$	$83.05_{(1.64)}$	$94.24_{(0.60)}$	$91.40_{(3.69)}$
	+Ours	$94.15_{(0.69)}$	$94.15_{(0.59)}$	$93.92_{(0.63)}$	$94.19_{(0.60)}$	$95.47_{(0.48)}$	$92.17_{(0.85)}$	$86.29_{(1.32)}$	$93.99_{(0.73)}$	$93.04_{(2.68)}$
	Δ	+0.75	+0.43	+0.50	+1.48	+2.53	+4.43	+3.25	-0.24	+1.64
BRACS-C (Bal. acc.)	Base	$62.59_{(5.18)}$	$62.41_{(5.13)}$	$65.99_{(4.21)}$	$61.82_{(4.71)}$	$44.70_{(5.34)}$	$54.89_{(5.38)}$	$60.28_{(5.02)}$	$57.61_{(5.19)}$	$58.79_{(6.18)}$
	+Ours	$72.78_{(4.79)}$	$70.29_{(4.90)}$	$70.70_{(4.97)}$	$72.60_{(4.67)}$	$51.40_{(5.14)}$	$58.42_{(5.49)}$	$65.31_{(5.48)}$	$59.87_{(5.14)}$	$65.17_{(7.36)}$
	Δ	+10.19	+7.88	+4.71	+10.78	+6.70	+3.53	+5.03	+2.26	+6.39
BRACS-F (Bal. acc.)	Base	$36.66_{(4.70)}$	$44.02_{(4.60)}$	$38.98_{(4.79)}$	$41.15_{(5.17)}$	$29.93_{(4.60)}$	$31.14_{(4.60)}$	$28.41_{(4.24)}$	$34.59_{(5.19)}$	$35.61_{(5.23)}$
	+Ours	$44.35_{(4.97)}$	$45.22_{(5.14)}$	$45.58_{(5.22)}$	$41.90_{(4.58)}$	$30.25_{(4.24)}$	$32.20_{(4.82)}$	$36.02_{(4.52)}$	$36.36_{(4.91)}$	$38.98_{(5.68)}$
	Δ	+7.69	+1.20	+6.60	+0.75	+0.31	+1.06	+7.61	+1.77	+3.37
EBRAINS-C (Bal. acc.)	Base	$90.77_{(1.92)}$	$89.86_{(1.97)}$	$87.97_{(2.02)}$	$88.11_{(2.04)}$	$84.74_{(2.39)}$	$86.97_{(2.21)}$	$75.91_{(2.35)}$	$83.24_{(2.36)}$	$85.95_{(4.44)}$
	+Ours	$91.86_{(1.84)}$	$92.15_{(1.79)}$	$90.41_{(1.94)}$	$91.74_{(1.68)}$	$86.17_{(2.25)}$	$90.36_{(2.02)}$	$83.02_{(2.26)}$	$90.34_{(2.04)}$	$89.51_{(3.02)}$
	Δ	+1.08	+2.28	+2.45	+3.64	+1.43	+3.39	+7.12	+7.10	+3.56
EBRAINS-F (Bal. acc.)	Base	$69.36_{(2.20)}$	$71.04_{(2.19)}$	$72.76_{(2.00)}$	$72.10_{(1.96)}$	$63.46_{(2.21)}$	$69.08_{(2.20)}$	$60.23_{(2.28)}$	$64.25_{(2.20)}$	$67.79_{(4.28)}$
	+Ours	$74.87_{(2.07)}$	$74.15_{(1.97)}$	$73.82_{(2.08)}$	$75.38_{(1.97)}$	$68.12_{(2.17)}$	$73.79_{(2.09)}$	$63.78_{(2.17)}$	$68.36_{(2.18)}$	$71.53_{(3.95)}$
	Δ	+5.51	+3.11	+1.06	+3.28	+4.66	+4.70	+3.54	+4.11	+3.75

G Computational Cost

MIST is designed as a lightweight plug-and-play module with minimal computational overhead. Figure 7 compares parameter counts before and after inserting MIST across eight MIL aggregators. With $K{=}8$ prototypes and rank $r{=}16$, the entire MIST module contains only 0.87M parameters, most of which (0.79M) belong to the shared projection layer replacing the baseline input projection. As a result, the net parameter increase is less than 0.09M for seven of the eight evaluated architectures. For example, ABMIL increases from 1.18M to 1.27M parameters ($+7.2\%$), while the overhead remains small for larger transformer-based models such as TransformerMIL (7.09M $\rightarrow$ 7.18M, $+1.2\%$) and TransMIL (2.93M $\rightarrow$ 3.02M, $+2.9\%$).

Interestingly, MIST reduces the parameter count of ILRA from 13.67M to 12.44M (-9.0%). This reduction arises because the shared projection compresses the input feature dimension from $1{,}536$ to 512 before the first attention block, reducing the size of subsequent projection layers. The efficiency of MIST further comes from the low-rank factorization of prototype-conditioned transforms, where each prototype contributes only a shared $\mathbf{V} \in \mathbb{R}^{D\times r}$ and a prototype-specific $\mathbf{U}_k \in \mathbb{R}^{r\times D}$ with small rank $r{=}16$, avoiding the cost of multiple full-rank transformations. Overall, these results show that MIST introduces spatial transcriptomic priors with minimal additional computational cost.

H Generalization Across Foundation Models

We repeat representative experiments with CONCH [11] as the frozen encoder, re-encoding the HEST patches and rebuilding the prototype bank in its feature space. Results are reported in Table 10.

I MAMMOTH with an H&E-only Control

Table 11 extends Table 6 with an H&E-only control, which replaces ST-derived prototypes with morphology-only clusters. Results on PIK3CA prediction in BRCA are reported in Table 8.

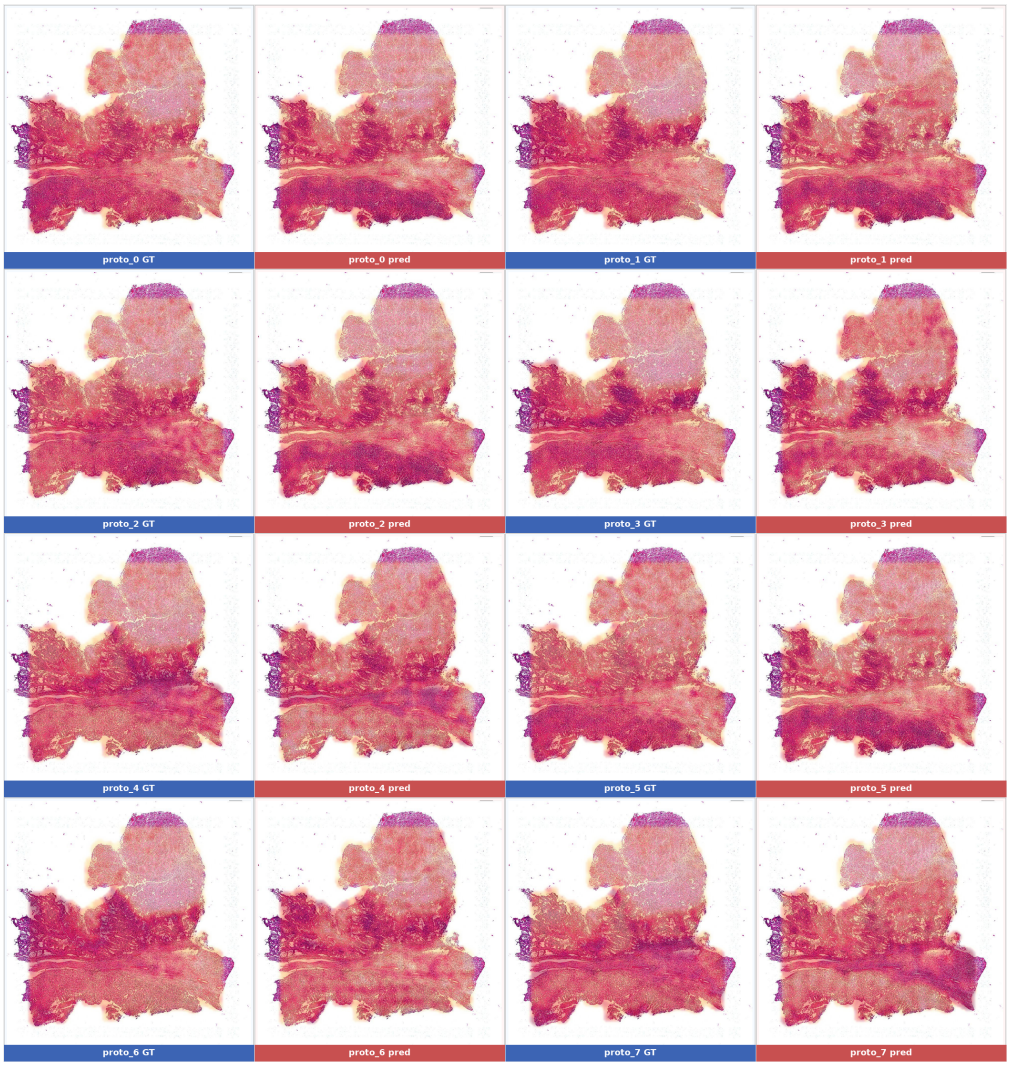


Figure 4: **Additional interpretability results on a held-out brain HEST slide.** Whole-slide prototype affinity visualizations generated from H&E images without transcriptomics at inference.

J Limitations

MIST relies on paired ST and H&E data for prototype construction, which limits the diversity of molecular concepts to the tissues and cohorts currently available in existing ST datasets. Although the learned prototype banks generalize across multiple downstream datasets and external cohorts, they may not fully capture rare tissue states, treatment-specific phenotypes, or underrepresented patient populations.

Choice of the projection layer. Injecting ST supervision at the projection layer avoids retraining the foundation model and keeps MIST compatible with existing MIL aggregators. Jointly training a pathology–ST foundation model may be more powerful, but requires substantially more data and compute.

Scope of the molecular contribution. MIST reorganizes frozen H&E features and cannot recover molecular variation that leaves no morphological trace; its contribution beyond morphology-only prototypes is modest but identifiable.

Encoder specificity and interpretability. The prototype bank must be rebuilt for each foundation model, and spatial interpretability is evaluated on one held-out HEST slide per organ.

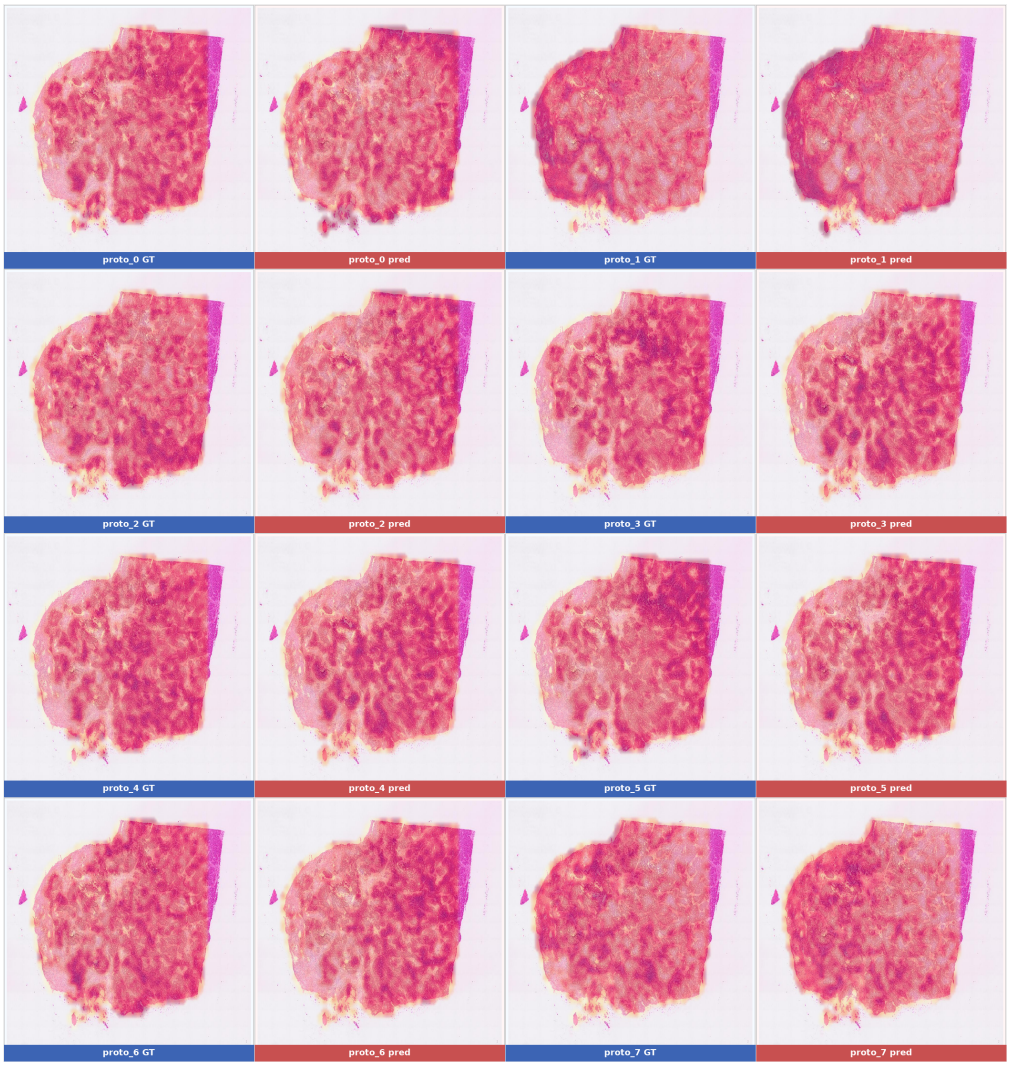


Figure 5: **Additional interpretability results on a held-out breast HEST slide.** Whole-slide prototype affinity visualizations generated from H&E images without transcriptomics at inference.

K Broader Impacts

MIST aims to improve whole-slide image analysis by introducing molecular priors into pathology foundation model pipelines without requiring transcriptomic measurements at inference. This may improve biomarker prediction, prognosis estimation, and tissue characterization while reducing dependence on expensive molecular assays. Because MIST operates as a lightweight plug-in module, it may also broaden access to molecularly-informed computational pathology systems in settings where transcriptomics is unavailable.

At the same time, MIST relies on molecular supervision derived from existing ST datasets, which remain limited in scale, tissue diversity, and population coverage. As a result, biases or underrepresentation in current ST datasets may propagate into downstream predictions and affect model generalization across institutions or patient populations.

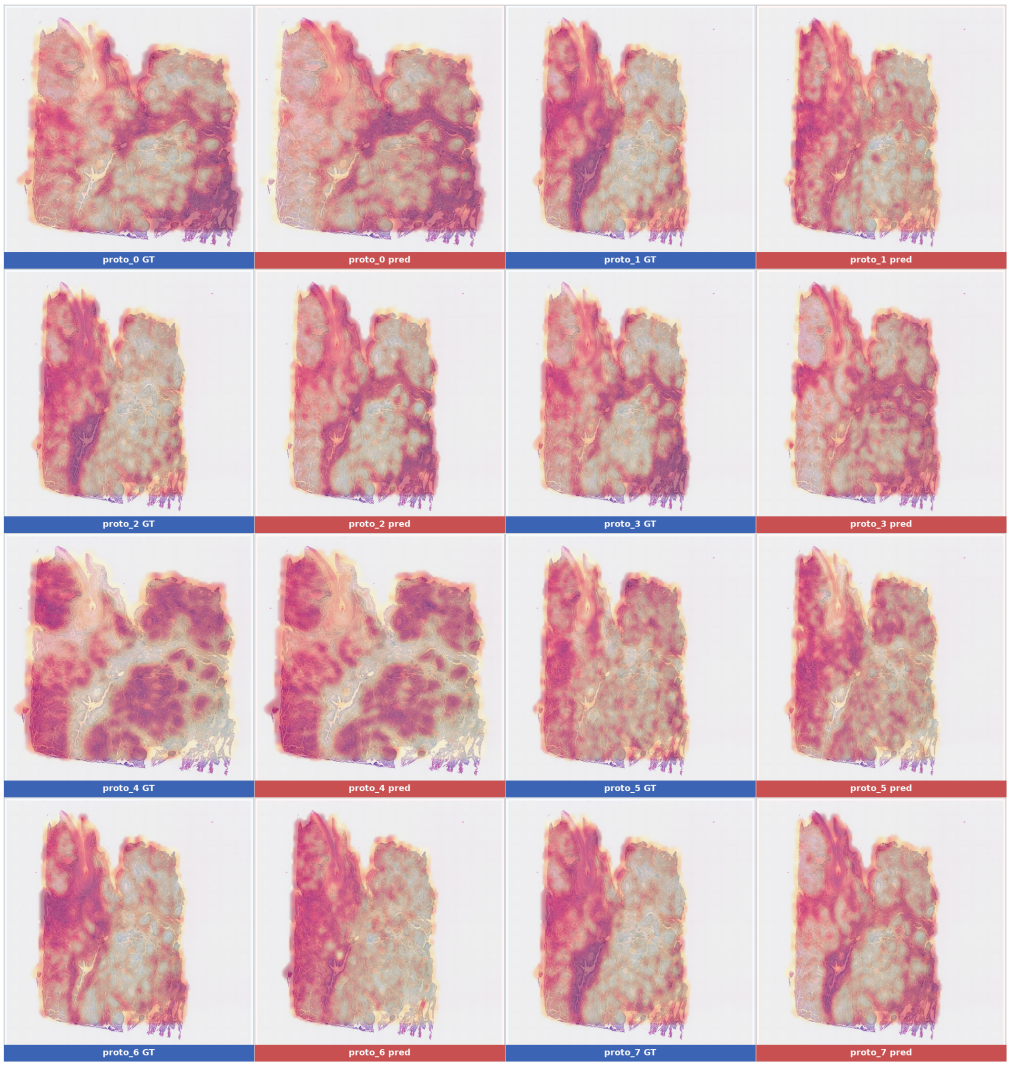


Figure 6: **Additional interpretability results on a held-out lung HEST slide.** Whole-slide prototype affinity visualizations generated from H&E images without transcriptomics at inference.

Table 10: **MIST with CONCH as the frozen encoder.** BRCA survival (C-index), BRCA biomarker prediction (mean over 4 tasks), and NSCLC subtyping (AUROC) using eight MIL aggregators.

Task	Status	ABMIL	CLAM	TransMIL	Trans.	ILRA	Mean	Max	DSMIL	Avg.
BRCA survival	Base	66.80	66.91	56.30	58.79	58.80	68.80	56.97	66.62	62.50
	+Ours	68.57	68.84	60.97	64.34	68.08	68.27	61.26	67.57	65.99
	Δ	+1.77	+1.94	+4.67	+5.55	+9.27	-0.53	+4.30	+0.96	+3.49
BRCA biomarker (4 tasks)	Base	74.93	74.88	72.68	73.61	70.60	71.98	66.07	72.41	72.15
	+Ours	75.94	76.22	74.42	75.53	72.72	73.56	67.87	73.92	73.77
	Δ	+1.01	+1.34	+1.74	+1.92	+2.12	+1.58	+1.80	+1.51	+1.63
NSCLC	Base	98.01	98.01	97.65	97.94	96.90	96.87	97.01	97.76	97.52
	+Ours	98.00	98.01	97.89	98.03	97.50	97.02	97.13	98.08	97.71
	Δ	-0.01	0.00	+0.24	+0.09	+0.59	+0.15	+0.12	+0.32	+0.19

Table 11: **MIST with MAMMOTH and an H&E-only control.** LUAD survival (c-index).

Task	Model	w/o MIST	+ H&E-only	+ MIST
LUAD	ABMIL + MAMMOTH	0.6324	0.6272	0.6647

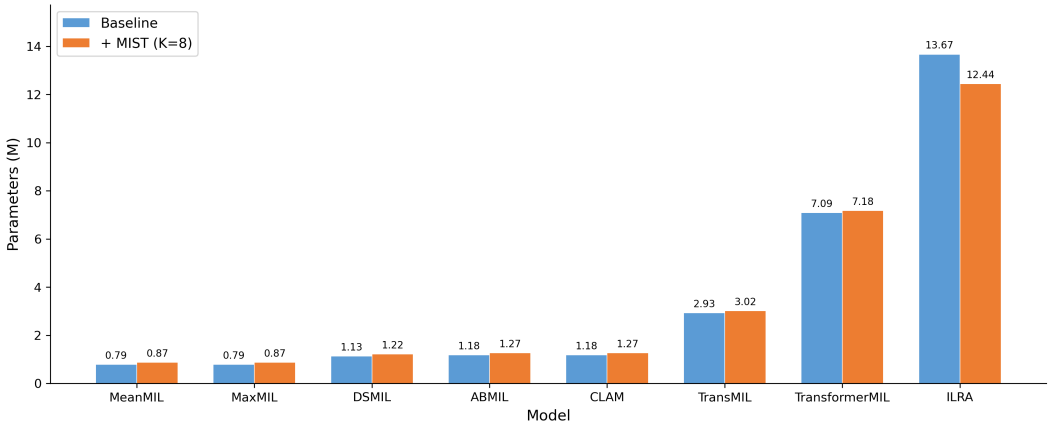


Figure 7: **Parameter comparison before and after inserting MIST.** Parameter counts (millions) across eight MIL aggregators with and without MIST (K=8, r=16). MIST introduces minimal overhead for most architectures and reduces the parameter count for ILRA through shared feature compression.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the paper's scope and contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses several limitations of MIST, including its reliance on paired spatial transcriptomics datasets for prototype construction and the limited diversity of current ST datasets. The paper also notes that rare tissue states and underrepresented patient populations may not be fully captured by the learned prototype banks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper states the assumptions underlying the theoretical analysis in the main text and provides complete proofs in the appendix. In particular, the decomposition showing how prototype-dependent routing can increase separation between morphologically similar but molecularly distinct patches is formally derived and proved in Appendix

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed implementation and experimental settings, including dataset descriptions, preprocessing procedures, model architectures, optimization settings, hyperparameters, evaluation metrics, and baseline configurations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code, prototype construction pipeline, and instructions to reproduce the main experimental results are publicly available at `https://github.com/YuchengXing99/MIST`. All datasets used are publicly available, and the paper provides detailed experimental settings, implementation details, and dataset descriptions.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the datasets, preprocessing pipeline, feature extraction procedure, MIL aggregators, optimization settings, learning schedules, hyperparameters, evaluation metrics, and training protocols used in all experiments. Additional implementation details and ablation settings are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports propagated standard errors in all main result tables. Survival, biomarker, and NSCLC tasks report standard errors across matched cross-validation folds; PANDA, BRACS, and EBRAINS report standard errors from 1000 bootstrap resamples of the test set. Results are consistently reported across multiple datasets, tasks, and MIL aggregators.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides the main implementation details required to reproduce the experiments, including preprocessing resolution, patch size, encoder architecture, optimization settings, batch size, training schedule, and model dimensions. The appendix additionally reports the parameter overhead introduced by MIST across different MIL aggregators.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors are not aware of any aspect of the work that violates the NeurIPS Code of Ethics. The research uses publicly available or appropriately licensed datasets and is intended to support computational pathology research rather than replace expert clinical decision-making.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential positive impacts of improving molecularly-informed computational pathology without requiring transcriptomics at inference, which may broaden accessibility in resource-limited settings. It also discusses potential negative impacts arising from biases and underrepresentation in current spatial transcriptomics datasets.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk generative models, scraped datasets, or systems with substantial misuse potential. The work focuses on computational pathology representation learning using existing research datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites the datasets, pathology foundation models, and baseline methods used throughout the experiments, including HEST, UNI2-h, and prior MIL architectures. The work uses publicly available research assets under their respective terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new public datasets, benchmarks, or software assets as part of the submission.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research conducted with human participants.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper uses existing de-identified research datasets and does not involve direct interaction with human participants or new human subject data collection.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large language models are not used as part of the core methodology, experiments, or scientific claims of this work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.